Housing Market Dissection:
A Data-Driven Analysis of the Housing Market in Dublin

Roshni Nath, B.E. Electronics and Telecommunications

Rajat Parihar, B.E. Electrical

A thesis submitted to University College Dublin in part fulfilment of the requirements of the degree of M.Sc. in Business Analytics

Michael Smurfit Graduate School of Business

August 2015

Supervisor(s):
Dr. Michael MacDonnell, UCD
Dr. Ernesto Diaz-Aviles, IBM Research

Head of School: Professor Ciarán Ó hÓgartaigh

Dedication

This work is dedicated to our parents, friends and well-wishers.

Table of Contents

List of Figures

List of Tables

List of Important Terms

Life of a Property: The duration of the property it remains on the site before it is taken off.

Views: Number of clicks on a particular property advertisement.

Min crawled date: The first date of its appearance on the site.

Max crawled date: The last date of its appearance on the site.

KDD: Knowledge Discovery in Database.

JSON: Java Script Object Notation.

CSV: Comma Separated Values.

Acknowledgements

Firstly, we would like to extend our heartfelt thanks and gratitude to our academic supervisor Dr. Michael MacDonnell for his immense support, guidance and presence. His time-to-time motivation and valuable suggestions helped us in making steady progress throughout the course of the practicum.

We would also like to extend our sincerest thanks to the IBM Research supervisors, Zubair Nabi for inception of the concept of the practicum, Dr. Ernesto Diaz-Aviles for taking this forward with his innovative ideas and constant guidance, his continued efforts to steer us in the right direction in terms of the practicum as well as introducing us to the other Phd. students Asmelash Teka, Aastha Nigam and Claudia Orella-Rodriguez working on the relevant topics under IBM and UCD provided us with enormous assistance and made things more approachable. We cannot thank them enough for their willingness to spare some time despite their busy schedule to conduct meetings frequently to discuss our work. Their support remains immeasurable.

Last but not the least; we would like to express a special word of thanks to our parents, friends and professors for being so considerate throughout the academic year. This project would not have been possible without their encouragement, patience and guidance throughout the work and is highly appreciated.

Abstract

Housing market is dynamic in nature. The volatility of housing prices acts as the catalyst for the growing concern among different sections of the people. This ambiguous market situation mainly due to rise in prices and shortage of supply hinders the sound decision making process as there are a number of unanswered questions in relation to the housing sector.

In order to address these questions, this practicum aims to dissect the housing rental market in Dublin. Ireland’s property website ‘www.Daft.ie’ is taken as the prime source of data and the trends of the properties are analyzed at large. Additionally, external sources such as Google Trends and Property Price Register are integrated to enhance the analysis. Tableau, a visualization tool is used to find meaningful patterns. A report in the form of dashboard is developed in Kibana to provide useful insights on a near real-time basis. In addition to this, a prediction model is also built that would predict the price of the property based on certain attributes. All this together would facilitate better decision making in terms of housing for all the concerned people.

Chapter 1: Executive Summary

1.1 Business Problem Definition

Housing market is a crucial sector, which contributes a lot to the national GDP and has immense influence on many other different fields. Because of its close-knit connection with the people, enterprises and other industries, understanding the market can significantly boost the decision-making capability of the individual.

Housing market has witnessed many highs and lows, even at present it continues to remain unpredictable with major issues of continued price rises and lack of supply of properties. In this work, Dublin housing rental market is dissected and an attempt is made to analyze the same by drawing key insights from it, which could be helpful for the stakeholders who are renters, property owners, real estates or even for the ones who are following the housing market in general. As the housing sector is dynamic and constantly changing, one should be updated with the current scenario of the market in order to participate in any form. At present, there are reports being published by other sources but those are issued on a quarterly basis, which are of not much help to a person looking to rent a house in a few days of time or for the real estate company who is planning on to place properties in the market soon. In this work, the objective is to provide a holistic view of the housing market by making it accessible in the form of report that can be generated on a real-time basis. Hence people can have a quick run over the trends prevailing in the market as to which area is being priced higher, which area is facing shortage, which has a better movement of property and many more.

For this purpose, data has been collected from a number of sources such as Google Trends, which would give the number of searches for any keyword and the largest property website in Ireland 'www.Daft.ie', as it is very popular among the people in Ireland and is considered as the most reliable source.

Considering the time constraint and the unwillingness of Daft to share any of their data, one month's data was collected from the daft website. Data set is from 2nd of July 2015 to the 3rd of Aug'2015 which we acknowledge is quite short span of a time to find some remarkable insightful data. Nevertheless, some interesting trends and patterns were observed, which would be discussed in detail in the result section that seem to be of interest and would definitely solve the purpose and cater to the aim of the practicum. In addition, the same setup of the practicum can be applied to the data of coming months in order to get the required results.

1.2 Assumptions

As mentioned, since there was no data prior to the start of our practicum, we managed to crawl one month’s data from Daft.ie website from 2nd of July 2015 to the 3rd of Aug’2015. The website is dynamic and every day they have about 250-300 new incoming properties in the rental section for whole of Ireland. The assumption is that for the data crawled on the first day i.e. 2nd July, the advertised date for the property would be the 2nd itself which means they are considered to have been first listed on 2nd of July irrespective of it being listed before the start of the practicum activity. It is so because Daft updates the advertisement date for most of its properties to the current date on a daily basis, hence for few data for 2nd July it wouldn’t have the exact arrival date (first listed date). Therefore, there are lot of properties, which would have their arrival date as 2nd July which actually is not true. This only affects the properties crawled on 2nd July and rest all are accurate. However, few properties would be from the month of May and June, which were still not taken up when the crawling was started and neither did Daft update their dates to current date. Likewise, for properties whose listing continue beyond 3rd Aug, their maximum listing date would be considered till 3rd of Aug only as the crawling was terminated from 3rd Aug. Therefore, the data would show many properties have moved out on 3rd Aug which would not be correct.

For other incoming properties crawled on dates apart from the 2nd of July and 3rd of Aug would be accurate because the exact dates were captured on when they arrived and dropped off the site as crawling continued on a daily basis.

1.3 Methods

Crawling and extraction techniques were used to retrieve the data from the website. The combination was essential, as the information required to fetch was split into two layers, one layer deeper than the other. Thus, the extractor would first return all the urls of the first layer with the help of the offsets given as inputs and then crawlers would return the exact information for each property from the second layer by making use of the urls returned by the extractor. Crawler would return the information in the way it was trained for the same. The output is extracted in Comma Separated Values (CSV file) and JavaScript Object Notation (JSON format). The CSV dataset was formatted and fed to tableau for analysis and the JSON data was indexed to Elasticsearch to visualize in Kibana (Elastic.co, 2015). Classification as a predictive modeling technique was incorporated in the form of a decision tree, which would predict the price of a particular property given the property type, advertiser type, area code, number of views, number of beds, number of baths and property life.

Additionally, social media was leveraged as part of the analysis. Twitter[1] was chosen, due to its growing means of communication which has a vast potential to uncover such communication networks and analyze the trend. However, the tweets were not analyzed in details, as the data collected was not in accordance to the requirement. Therefore, the count of tweets were collected as per the text search during the course of the analysis. Another social media accessed was LinkedIn[2] to gather the information on the job posting. However, due to few limitations which are mentioned in details in the methodology section, the desired data could not be collected. Alternatively, Google Trends[3] was used to see the search trends for the jobs in Dublin if it has any impact on the overall housing demand. Data from property price register[4] was also collected to get the details of all the properties sold for the month of July 2015 to understand the rent/sell equation of the area. All these data were incorporated in the data analysis.

[1] https://www.twitter.com

[2] https://www.linkedin.com

[3] http://www.google.ie/trends

[4] https://www.propertypriceregister.ie

1.4 Output

As the housing sector is divided into three broad categories, results are tailored to meet the demands for each of them. With respect to **Real Estate Agencies**, this practicum helped them to forecast the housing demand in the market. The property views and its life are considered as the measure for the demand and are correlated with other attributes such as the count and the price, to forecast the demand of the property as per the property types. On the basis of the analysis and the trends observed which are discussed in details in the results section, it is suggested that more Apartments can be built in Dublin 5, Houses in Dublin 17, Flats in Dublin 7 and 24 and Studios in Dublin 13. In addition, the predictive model that predicts the price can be beneficial for the real estate as they now know which combination of attributes that includes the number of beds, apartment type, area etc. would fetch them the price they are looking for. From the view point of **Property Owners/Investors**, they can access the Kibana report, which is in the form of a near real-time dashboard to get useful insights on the views and life of the properties in the areas, which as discussed is considered a measure of the demand. Hence, the Property Owners/Investors can plan their properties accordingly. Also, the data from the property price register would show them which area yields more on the basis of the rent/sell ratio. The predictive model lets them decide on the price they should let in to the tenants. For the **Renters/Buyers**, same report would be very convenient to use during their decision making phase. As the reports consists of heat-maps, line charts and bar plots, they are very easy to understand and easily interpretable for the common public. It gives insights on the prices across the Dublin area, also provides an idea of the properties available in these areas. Interested renters or buyers can see which area suits their budget on the basis of the number of beds and the property type as they like. The predictive model would be an add-on benefit that could let them know the price of the property they like as per their choice of the attributes. In addition, the Google Trends can show the relation of the job search and the movement of the properties. Therefore, the practicum successfully works towards meeting the demands of the stakeholders and adds good value to their decision making process.

Chapter 2: Introduction

2.1 Opening Remarks

Housing is far beyond just bricks and mortars. One needs to consider a lot of factors when it comes to either buying a house, renting a housing or letting a house. Few among the many factors usually considered are the price, area, type of the property etc. Among all, price plays the most important role. Price has been ever increasing. Demand for houses is soaring where supply still lags behind. Ronan Lyons, in-house economist of Daft.ie says Ireland has a growing population but not a growing housing stock, which is putting pressure on the housing and rental market (Pollak, 2014). Given this situation, it is really necessary to dissect the housing market in order to understand the challenges and action what is best.

For this purpose, we divide the housing market broadly into 3 major stakeholders.

1) Real Estate and the property domain at large: Their challenge is to understand the market and size it accordingly.

2) Property Owners/ Investors: Their challenge is to identify areas to improve their dividends (simplify the rent/sell equation).

3) Prospective Renters/ Buyers: Their challenge is to decide a property of their choice to rent/ buy.

In our work, we would cater to the needs of each of the above stakeholders by publishing a report that would be in the form of a dashboard, which would act as a reference guide and assist in their respective area of decision-making. In this project we delve into relevant literature, identify fit to purpose data source, tools, techniques and algorithms, best suitable for our objective of the project.

2.2 Business Context

In this section, we will walk through the background of the housing market and its current situation before entering the main context.

As it is experienced housing market remains vague in many aspects due to its volatile nature. It has seen several changes since the last few years. Ireland has seen a massive property price boom in 1997 to 2007 with average used house prices up by 268% and prices of new homes up by 216% (Guide, 2015). Ireland's decade long house price increase was one of the longest and biggest in Europe, it was mainly fueled by strong economic growth, immigration and generous tax incentives and grants from the government. The situation turned around in 2006, 2007 when interest rates hikes left many borrowers in difficulty, causing a housing market crash. The bubble busted in 2008, the world's biggest property burst. In 2008, the housing price index fell 12.4% compared to 2007. In 2009, it further fell to 18.6%. Further declined to 10.5% in 2010. In 2011, it increased to 16.7% again in 2012 dropped to 4.5% as compared to the previous year. The housing market started to recover in 2013 with prices rising by 6.4%, which was mainly the result of economic growth. The economy expanded by 3.6% in 2014 and is projected to grow by 3% this year (Guide, 2015).

Rental Market

The rental market has seen a significant increase in demand in 2014. Dublin region experienced the highest increasing rent in 2014 largely due to improved economic conditions and increase in employment in Dublin. However year on year rents in Dublin has eased from 16% to 6% since April 2014. Rent for the houses has increased by 6.5% and apartment by 7.8% compared with quarter 1 of 2014 (Lyons, 2015).

According to Simon Stokes (Founder of Stokes Property Consultants Ltd. Dublin) increasing demand of the residential market owing to the shortage of supply consequently led to the increase in the rental prices which in some locations are 5 to 10% below the peak of 2007.

Simon stokes anticipate the overall increase in rental rates in 2015 along with about 30% increase in rents in Dublin's prime areas like Southern City, Docklands due to good connectivity and close proximity to city centre (O'Donovan, 2015).

According to CSO index Dublin prices fell by 1.9% in Q1 2015, but increased by 1.0% in April and then fell by 0.1% in May. Prices in the capital are growing at a faster rate. Asking prices have seen a growth by almost 5% in the first few months of 2015. Dublin South City continues to lead price gains with median asking prices still up by 25.3% in the year 2015 (Mac Coille, 2015). Dublin south faced median asking prices up by 13.6% in the year to Q2. While Dublin west saw a 12.3% annual increase. North Dublin price were up by just 3.2% a year which is the slowest recovery of all in terms of asking prices (Lyons, 2015).

The average cost of renting in Dublin stands at €1,142 per month, up 5.5 percent - the highest increase in the country (Independent.ie, 2015). Some 113,000 homes are rented, almost 40 percent of the national stock. Some 2600 new home have been built in the initial three months of 2015, approximately 25% higher than the same period in 2014. Dublin population could grow up to 100,000 during the 2010's, contrastingly fewer than 10,000 new homes have been built in the capital (Lyons, 2015). In short the problem is too few homes. These problem is evident in Irish Society from working, homeless to the students who are being intimidated by expensive housing (Uğur, 2015). In around May 'construction 2020' a strategy for a renewed construction sector was published which strategized to restore the construction industry, also address housing and employment requirement nationally. As mentioned, increase in demand for properties followed by the shortage in supply is one of the key concerns faced by the market in 2015 (Independent.ie, 2014). SCSI members reveal that there are many hurdles in the path of the construction sector revival like the high construction costs, stringent building regulations. As per Daft report, the trend of list prices are similar to that of the transaction prices. Dublin price inflation has cooled from 25% in autumn in 2014 to 17% in early 2015.

The combination of a range of factors are providing the perfect storm of conditions for continued runaway house price inflation. On top of the chronic supply issue, the banks are now starting to lend again at very low interest rates. While a return to mortgage lending is a good thing for homebuyers, it can only increase house price inflation further in the capital (Independent.ie, 2014).

Therefore, we chose to analyze the rental market in Dublin, which constitutes the majority of the issues. As we have divided the rental market in three stakeholders, we strive to meet their needs in this practicum by creating a real-time dashboard. For **Renters**, we focused on providing an overall glance of the market situation to the people concerned and enable them to take their best foot forward. We recommend the **Real Estate** sector which area to focus on to ease on the supply demand issue. Also, for the **Investors/Property Owners**, we give them an idea of the rent/sell ratio across the areas in Dublin. For all this purposes, we used combination of the extraction and crawling technique as used by Rys,M (Rys and Fai Yau, 1997). We have incorporated lot of visualization to help grasp the info in lesser time. Data visualization can simultaneously examine several aspects of data sets enabling to discover rich analytical sights that might not have ever surfaced (FEW, 2007). Also for the real estate data being classical spatial temporal, data visualization helps to draw previously unknown insights from them (Sun, 2014). Classification has been used to draw customer preferences in real estate data (Gupta and Dubey, 2012). Likewise, in our work we used classification technique to build a prediction model designed on selected area code that could cite the price for the property of the mentioned attributes. Social media sites such as twitter were used for our analysis, as it not only enables networking capability but also enable to understand the thoughts on a particular topic (Needtagger.com, 2015). External data source like Google Trend has been used. A lot of prediction has been done in various fields on the basis of Google Trends, Ettredge et al. (2005) published the initial paper that revealed that web search data is useful in forecasting economic statistics, which examined the US unemployment rate (Choi and Varian, 2011). In our work, we use Google Trends to map the job searches with the movement of the property in the housing market.

2.3 Aim/ Business Question

The practicum aims to meet the challenge of each of the stakeholders. The insights would be presented in a dashboard that would act as a report, which can be accessed on a near real-time basis. It would be helpful in getting an overall trend of the market at that point of time and the needed decision can be taken by the respective stakeholders accordingly. The report would be in consonance with the following questions with regards to each stakeholder:

1) Real Estate and the property domain at large: Can we help them forecast the housing demand for particular areas?

2) Property Owners/ Investors: Can we simplify the rent/sell equation for them?

3) Prospective Renters/ Buyers: Can we empower them to make a more calculated decision by giving them a holistic view of the market?

2.4 Outline

This section details the flow of the practicum. It mentions each section and the gist of it as they developed during the course of our work.

- Literature Review

In chapter 3, we mention about the in-depth study of the literature and the work done in areas of Knowledge Discovery in Databases (KDD), Crawling, Classification, Data Visualization, and Social Network. It justifies the reasons behind the selection of the following techniques:

KDD: Since we had to start from the scratch of selecting the data source to finding useful insights, it follows the knowledge discovery process.

Crawling: Since no data set was provided only way was to crawl Daft site to get the data. Given the structure of the site and based on the literatures read we decided to use a combination of an extractor and crawler which best suit our purpose and nature of the site.

Classification Techniques: Decision Tree seemed most suitable to build the prediction model.

Data Visualization: To understand how to make the best use of the crawled data with the help of visualization tools with respect to the property market.

Social Networks: To understand how external sources especially Twitter and LinkedIn and Google Trend can be incorporated to analyze the housing market.

- Methodology

A detailed discussion of the methods implemented and the tools and techniques used, in accordance with the step by step process of KDD which involves data selection, data preprocessing, data transformation, data mining, data interpretation and evaluation are discussed in chapter 4.

- Results

Chapter 5 consists of the output obtained from the above methods, trends, patterns and the rules found. The graphs and the charts that we found of most importance and relevance to the business question of the stakeholders were collaborated in the dashboards of Kibana, which would be available in the form of near real-time reports.

- Analysis and Discussion

In chapter 6, the rules and patterns were interpreted with respect to the business questions of the stakeholders.

- Measure of Success

Chapter 7 mentions the feedbacks and the statistical validations conducted for computing the success of the work.

- Learnings

Chapter 8 mentions the learnings we have undergone during the course of our work.

- Conclusion

Chapter 9 wraps up the work that we have done in Dublin housing market domain.

- Future Work

Lastly, chapter 10 discusses the work that can be done in future and also the work which we could not implement due to some valid constraints.

2.5 Contribution

To dissect the housing market, we collected data from various sources like Daft, Twitter, LinkedIn, Google Trends. However, the volume of data obtained differs for each of these sources. Majority of the data was taken from Daft. These data were visualized and analyzed and various trends and patterns were observed

The dashboard that is produced with the help of the collected data has all the important insights pertaining to the stakeholders. Also, the prediction model in the form of a decision tree helps to predict the prices when given the various attributes of the properties. This would be helpful for all the stakeholders. For instance, if the renter wants to rent a house he would have a fair idea of the price that would be charged. Likewise, the real estate can price their houses according to the model, which would be as per market standards.

Chapter 3: Literature Review

3.1 Introduction

In recent years, there has been an intensified need for data and analytics to be implemented in the corporate decision-making. Both public and private sector have begun to rely on analytics to generate data driven insights that would enable better business planning. Business intelligence, data-ware housing, big data are the means to an end, which is to gain insights that would be of real value.

With the help of analytics real estates are able to identify the market trends in housing sector, establish the happenings in particular areas, able to predict particular changes in certain geographical areas and determine how the business or the public in general should react to it. Foremost in this sector or any competitive sector is to understand the need of the customer what they need, when they need, which is done by building better customer relationships with them through having real-time customer insights. Ensuring that there is proper access to the real-time data by implementing relevant services. Focus has been made on streamlining the business activity by generating customer branded reports and specific targeted campaigns. Efforts have been taken to observe and do a comparative study of the trends for different areas (or between specific areas). The study is done taking into account the measure like the average price, listings volume, sale volume, rental yields, time on market, vendor discounting and time based changes in median sales price (Corelogic.com.au, 2015). Some residential property agencies have gone a step forward by extracting customized data related to the housing sector. Typical customized data extracts includes attributes such as its price, number of bedrooms/bathrooms, roof and wall type, garden area, fireplace, gymnasium, pools, features such as amount of property in that area, air conditioning and many more. Analytics have been used to gain insights of the market like number of property sold, time to sell, and current trends. New listings are tracked, fair market value of a property is established and presented in the form of reports. Analytics in real estate exposes the stakeholders with as much information as possible on properties, streets, suburbs and postal codes. Customers are segmented and then targeted to enhance customer acquisition, retention, and cross sell opportunities.

A lot have been done in descriptive reporting there is lack in using data to present future scenario in a clearer way. There is not much done is providing some actionable recommendation based on these data.

The past decade have brought along tremendous pressure on businesses and respective stakeholders to optimize their decisions. Businesses need to optimize their performance, comply with demands, deliver to customer efficiently and identify new areas for increasing revenues. The same goes with decision makers, they need to have a clearer picture of what they are venturing upon.

By applying analytics, hidden correlations and patterns can be detected, predict future trends, and deliver previously unknown or inaccessible insights. In the housing sector, real estate are using predictive modeling and various data mining techniques to generate models that can predict future scenarios. Advanced data visualization has been adopted which permits a more interactive display of pictures and charts instead of plain rows and columns and statistics. Advanced visualization help understand complex data and increases the reach and influence. However not much has been down in prescriptive modeling, simulation and optimization. It not only should be able to predict the future but also recommend actions to the decision makers.

As the data in real estate market is enormous, it is extremely necessary to make use of powerful process for analysis of such data and drawing interesting insights from it. Data mining is one such process of inferring previously unknown knowledge from such large data. In the recent times, data mining has become one of the most popular tool for extraction of unknown data and establishing patterns to assist in decision-making. In the paper ‘identifying customer interest in real estate’ the authors Vishal Venkat and Swapnil Vijay have also made use of data mining techniques to predict most suitable area for customer (Venkat Raman, Vijay and Banu K, 2014). Weiss et al. however, divide DM into two categories: prediction (classification, regression, and times series) and knowledge discovery (deviation detection database segmentation, clustering, association rules, summarization, visualization, and text mining) (Chukwugozie Nsofor, 2006).

In our practicum, classification is implemented as part of the predictive data mining method and visualization under knowledge discovery. Knowledge Discovery in Databases (KDD) is an umbrella name for all those methods that aim to discover relationships and regularity among the observed data (Fayyad et al, 1996) (Chukwugozie Nsofor, 2006). KDD composes of a number of steps, from identification of initial business aims to the application decision rules. The stages of KDD are data collection, pre-processing, data transformation, data mining, which are discussed in relation to our practicum detail in the later section.

3.2 Predictive Modeling/ Classification

Classification technique has been increasingly used in the real estate sector to identify the customer behavior and preference (Gupta, A. and Dubey, G. June 2012). In our practicum, the method for classification used is decision tree.

Recent development in information technology has significantly advanced the generation and consumption of data in our daily life. As a consequence, challenges such as growing data centers, the need of intelligent data analysis and the scalability, reliability for large or continuous data volumes are now moving to the desktop of business managers, industry experts or even end users. The KDD and data mining backed up by disciplines such as artificial intelligence, machine learning and statistics is designed to explore and search the useful information from the big data sets and helps the business managers to make better decision and increase profit (Maimon and Rokach, 2010) (Fayyad et al, 1996). One of the approach of Data Mining under predictive analytics is classification, which is defined as a task of assigning objects to one of several predefined categories (Frawley et al, 1991). It is a pervasive problem that encompasses many diverse business analytics applications. Such as in business intelligence, data classification has close ties to data clustering, but where data clustering is descriptive, data classification is predictive (Golfarelli and Rizzi, 2009). In essence, classification consists of using variables with known values to predict the unknown or future values of other variables. It can be used in direct marketing, real estate sector, insurance fraud detection or medical diagnosis (Kimball, R. et al., 2008).

Random forest technique was implemented to build the tree. Breiman (2001) proposed random forests, in which the node is splitted using the best among the subset of predictor randomly chosen at that node (Liaw and Wiener, 2002), contrastingly in the standard trees the split is done by choosing the best split among all variables. Random forest appeared to perform well compared to other classifiers and is known to be robust against overfitting (Liaw and Wiener, 2002). It uses just two parameters, which are the number of trees to grow in the forest and the number of variables in the random subset at each node. Hence, it is considered to be user-friendly and it's not so sensitive to the values.

Due to the stochastic behavior of the real estate sector, classification methods were used in this practicum to build a predictive model that forecasts the property rental prices.

3.3 Crawling

The World Wide Web have grown immensely from a few thousands of pages in 1993 to more than a billion pages at the present time. These websites offer large expanse of information. It lies in heart of any business. It is essential to locate the important information and make the most out of it. In order to access these vast amount of data, an extraction or transformation tool is required to transform the information in a form which renders both effective and efficient data extraction and information integration.(Rys and Fai Yau, 1997)

There has been remarkable work on extraction and structuring data from web sites using wrapping and filtering techniques where information can be extracted from the predetermined set of web pages, also crawling can be done through the web to discover some relevant information. Search engines rely on collection of web pages that are acquired with the help of web crawlers (Shkapenyuk and Suel, 2001). Crawlers (also known as robot or spider) are basically a program that visit the entire websites or reads their specific pages. Crawlers crawl through a website page wise following every link to the other pages on it until all the pages have been crawled. Roughly it starts off with a set of url u_0 which is placed in the queue, where all the urls to be extracted are kept and prioritized. From this queue, the crawler gets a URL, download the page, retrieves any url in the downloaded page and places the new url in the queue. This process is repeated until the crawler is stopped (Cho, 2001).

Web crawling helps in getting the data that would help in making business decisions. Web crawlers are as old as the web itself. Mosaic and Matthew implemented the World Wide Web Wanderer and was used until 1996 (Olston and Najork, 2015). However, design of these early generation crawlers did not handle scalability issues. Few years' later crawlers started being widely used in search engines Alta Vista, InfoSeek, Excite, and Lycos (Shkapenyuk and Suel, 2001). Google has made the web crawling technique popular by using it in their search. Brian and Page's 1998 paper outlines the architecture of the first generation Google crawler. They were among the first to realize the importance of data on the web which remained untapped. At present they make use of thousands of crawlers in the web and index everything they can possibly find. Recently, Yan et al. described IRLbot, a single process web crawler that is able to scale to extremely large web collections without degrading the performance (Shoaib, Farooqui and Zunnun Khan, 2015). Most commercial web crawlers simply index a web page for further retrieval. They might extract some data for the crawling purpose but it's not user specified. On the other hand, the standard wrapper of web sources usually extract data from a predetermined set of pages. In such cases, one need to know the hyperlink structure of the sources and the extraction procedure for every document encountered.

Also, many a times the information on the web is scattered among many different pages which hurdles the automated extraction process the condition becomes more complicated when the different pages are generated dynamically or pages are added and deleted. This is when both extraction and crawling is needed (Rys and Fai Yau, 1997).Also the process has to be scalable so that it can extract large amount of data from larger web sites. Since the site that has been used in this work as the source of data is dynamic and has information split over two pages an extraction tool is required which is suitable for navigational browsing as well as dynamic nature of the site which is basically the integration of crawling and extraction techniques. However, most of the processes do not include both extraction and crawling technique.

In order to solve such problems combination of the two approaches needs to be implemented, crawler which crawls through the portion of the website and an extractor then extracts the information in a specified way required for further processing. For this practicum, similar concept was approached of combining both extractor and a crawler. Some other methods have been studied in relation to the extraction of information from multiple pages. As per Michael and Fa, in Ariadne one needs to create a domain model of the data to be extracted and associate wrappers to the domain concept (Rys and Fai Yau, 1997). These wrappers are specific to the sets of equivalent pages, work as query interface to be used by query planner to compose the final query plan. It allows to extract data however the domain model need to exactly reflect the structure of the website. In the approach used in this practicum, the structure of the website is discovered while crawling and extraction of the data like the one described by Michael Rys Ka Fai Yau. Adridne wrappers are produced through machine learning techniques whereas the tool used for this work, the combination of crawler and extractor uses a template based extraction mechanism. Hence, it provides more adaptability and is dynamic in adapting to different sites. Akira also suggested another method for extraction that as a web only information integrator and query engine is more closely related to information integrators. Though it combines crawler and extractor, it is relevant for our purpose. However, it does not use web pages as a logical unit instead uses fragments. The suitability of the queried data received from the web is decided by IR technique which makes the extraction and the crawling less precise compared to the method that is chosen to implement which is more suited for querying free text pages and semi structured documents.

3.3.1 Architecture of the crawler

The extractor crawler combination tool extracts data from a set of hyperlinked pages on the web to be used for further processing. A specification file and set of extraction template is provided as the input. The former identifies the page of interest and mapping between these pages and the extraction template to be used is specified. The set of extraction template makes use of pattern matching rule to identify, extract, and transform the required data from a web page. The architecture of the extraction crawler consists of the following parts:

Web access manager: The extractor retrieves the page from the web through the web access manager and not directly through the web in order to avoid multiple retrievals of the same page.

Extractor: It performs an individual extraction on the given web page based on the specification given by the extraction template. The extraction template specifies the extraction and the transformation process using simple pattern matching rules as below:

[source object, pattern1, object1, pattern2, object2, . . .], [object1 (or function(object1)), pattern11, object11, . . .], . . .

The first rule applies pattern1 to the text from the source object and stores part of the data in object 1.In case the pattern doesn't match the next pattern is applied until one matches. The second case applies pattern 11 to object1 and extracts data to object 11.

Reassembler: It reassembles the OEM object graph generated by the extractor into the resulting OEM graph.

Controller: It controls the operation of the extractor and the reassemble according to the user specification

Loader: The final crawling result is either post processed by a wrapper or it can be passed to the loader that loads into the specified database.

3.3.2 The crawl extraction specification

It consists of the following main parts: start url, depth of the crawl, inclusion and exclusion list, the document map, sleep. The start url specifies the root of the crawl from where it has to begin and follow the hyperlinks in it recursively until the page depth specified has been reached. Within this domain is the inclusion and the exclusion lists which mentions where all to crawl and where not to crawl. The extraction crawler allows to provide the set of included and excluded web page by MIME type for instance include only text/html or by passing a set of required prefixes or suffixes for the url. The document map assigns the selected template to the selected pages. It groups the pages according to their URL, and associates a list of extraction templates with each group.

3.3.3 The crawling process

The controller starts the crawling process by identifying the document map for the first group for which the start url matches the group criterion. If there is a match, extraction are attempted using the associated templates in the specified order until the first successful extraction or if none of the templates results in a successful extraction the controller continues to search for another matching group for the url. This process continues until a successful extraction is searched or until the templates in all the matching group for the url have been tried. Once the extraction is successful, the re assembler receives the extracted data in the form of an OEM graph and the controller receives a list of all the urls to which the crawl page refers. The controller repeats the process until the max page depth is reached.

The above mentioned method allows selective crawling of a website with only knowing the abstract structure of the website (Rys and Fai Yau, 1997). The only minimal input to be provided are the starting URL and mapping between specific document and template is required. In addition, the abstract view of the extraction crawler is flexible enough to cope up with the changes in the site structure.

3.4 Data visualization

Visual analytics has become an important tool for gaining insight on large and complex data. Many statistical tools and data transformations such as projections, binning, and clustering have been coupled with visualization to help understand data better and quicker. Visual analytics facilitate the process of data understanding by means of interactive visual metaphors. Dolfing describes the visual analytic process as a series of transformation that facilitate insight from a collection of heterogeneous data sources. The transformation can be categorized as data/visual transformation, which draws representation from increasing visual mappings and structure, which converts these into visual elements used by visualization interface (Correa, Chan and Ma, 2015). Multivariate analysis is the core of visual analytics. Methods related to it are regression, generalized additive models and response surface analysis. These methods find the relationships among variables to fit models to multivariate data. Yang et al integrate analysis tools with visual exploration of multivariate data using nugget management system. Barlowe et al introduce the derivatives of dependent variables to find correlation between them (Correa, Chan and Ma, 2015).

Real estate market data are classical spatial temporal data, indicating geographical distribution of houses, trends in house price, sales volume, and other unknown trends. Visualization helps present, analyze and discover the hidden stories intuitively, efficiently and interactively (Sun, 2014). Visualization with the help of images has been immensely appreciated for the benefits it is offering to business. It provides a powerful means to both understand data and communicate it to the audience at large. Businesses are increasingly shifting towards visualization based data discovery tools. It not only allows graphical representation of data but also facilitates altering the nature of the display, filtering out according to its relevance, and drill into deeper level of detail and have subsets of data across multiple graphs. These instigates better response of the viewer, resulting in insights far better than the traditional approaches. Edward Tufte popularized graph names small multiples which consists of series of small graphs arranged together so that they can be compared easily (FEW, 2007). For instance, multiple types of graphs like the bar, line and scatter plot could be plotted together which would result in examining several aspects of the data, realizing hidden connection which could not have been visible is viewed separately. Dashboards are one single platform that combines all the displays that is needed for efficient decision-making. Another very important aspect of data visualization is the geo spatial visualization. The popularity of Google earth and other similar websites have contributed a great deal to this interest. Much of the information required in case of our project is tied to the geographical locations for instance the sale, price, and type of property available in which area. By making use of the codes and fetching the longitude and latitude it is able to provide with a better visual display experience each day.

The primary challenge in big data is its volume variety and velocity. The rapid generation of the data can lead to significant outcome only if data can be analysed quickly in hours rather than months. It is extremely vital to reduce the latency between data capture and to action it, it is where visualization comes into picture. As the real estate sector is constantly changing, data visualization is what seemed appropriate for our project. Output in this sector is mainly in the form of reports. Hence, better the visualization better is the data communicated.

3.5 Social Network

Social media has been gaining popularity not only for its social presence and its networking capability but how it is helping businesses in their marketing, advertisements, promotions and likewise. It helps in targeting the specific set of customers and focus on them. It is an easy and efficient way to learn about the audience. It enables us to understand the needs or requirement of the masses as a whole. We therefore made use of twitter and LinkedIn to gain insights with regards to renting among the people in Dublin.

3.5.1 Twitter

Twitter is one among the most popular way of expressing thoughts on a particular topic. Twitter helps to extend the reach to a larger section of the people and generate word of mouth. It provides an idea of what is trending what are their views and their area of interest. Many businesses have incorporated twitter as a means to grow their business. The finding from "Small Business Customer insight study" (2014) show that followers can help achieve reach, sales and word of mouth (Needtagger.com, 2015). Hence, twitter was used as part of the practicum to gain insights on the housing market of Dublin by collecting tweets related to housing/renting in Dublin made by the public and also the real estate companies to analyze if that has any effect on the housing market.

3.5.2 Google Trends

In this work, we incorporate Google Trends, which provides real-time volume of the queries that users enter in Google. These query indices are correlated with economic indicators and could be useful for short-term predictions (Choi and Varian, 2011). It certainly does not predict the future much but it helps in predicting the present. An initial paper was published that revealed that web search data is useful in forecasting economic statistics, which examined the US unemployment rate (Ettredge et al., 2005). In economics, Google search insights can be used to predict several economic metrics like unemployment, automobile demand etc. (Choi and Varian, 2011). Recently, Google search data has been used to examine how job search responded to extensions of unemployment payments (Baker and Fradkin, 2011). The predictability of Google Trends data itself has been described, pointing out that a substantial amount of search terms are highly predictable using simple seasonal decomposition methods (Shimshoni et al., 2009). Goel et al. [2010] highlights some of the limitations of web search data. As they point out, search data is easy to acquire and is often helpful in making forecasts, but may not provide dramatic increases in predictability (Choi and Varian, 2011). Although we agree with this, we typically find significance if not forecasting. Therefore, we used Google Trends to analyze the effect on the movement of the property based on the volume of the job searches.

Chapter 4: Methodology

4.1 Overview

In order to find the useful insights of Dublin House Rental Market, we have collected data from various sources such as Daft.ie, Twitter, LinkedIn, Google Trends and other property sites. The methodology is based on the KDD process and it involves steps like Data Selection, Data Preprocessing, Data transformation and Data mining for the extraction of useful trends and patterns from the large dataset. The classification is used as a predictive modeling technique to make a Decision Tree.

4.2 Tools and Softwares

The tools, programming languages, softwares and other resources used are:

- WebCrawler is used for web crawling.
- Python, MS Excel and MS Access are used for data preprocessing, data cleaning and data parsing.
- Amazon Web Services[1] and Google Drive are used for cloud storage and collaboration of data and documents.
- Github[2] is used as the code and dataset repository.
- ElasticSearch (v1.7.1) is used for data indexing and Kibana (v4) is used for creating a near real-time dashboard.
- A command line tool 'Stream2es' is used for adding data into Elasticsearch.
- Google Maps API is used to retrieve Geocodes (Coordinates).
- Google Trends is used for finding the correlation between jobs and rental market.
- Tableau (v9) is used for data visualization.
- R studio (v 0.99) is used for creating a predictive model.
- Complete experiment has been performed on Windows 7 x64 and MacOSX (Yosemite) platform.

[1] https://aws.amazon.com

[2] https://github.com/

4.3 Brief description on Elasticsearch and Kibana

As suggested by IBM research, in order to create a user friendly near real-time dashboard with an option to update the visualizations, actionable insights and patterns in real-time on daily basis, we have decided to go with Elasticsearch and Kibana tools driven by the open-source vendor ‘Elastic’. Main attractive feature which made us take up Elasticsearch as our choice is its ability to handle real-time data, since data is fast moving with this feature we wouldn’t have to wait in order to draw insight from it, we can immediately start search and analytics.

Elasticsearch is an enterprise solution for the data extraction problem. It has the benefits of real-time data and real-time analytics. It is a distributed, open source search and analytics engine, designed for horizontal scalability, reliability, and easy management. It allows multiple indices to be queried independently or in combination. It uses Apache Lucene full text-search capabilities, which is again an open source product and combines its speed of search with the power of analytics via a sophisticated, developer-friendly query language covering structured, unstructured, and time-series data (Elastic.co, 2015).

Kibana is an open-source data visualization platform that allows the interaction with the data through stunning and powerful graphics. It is basically a log-data dashboard which uses the indexed data of Elasticsearch in JSON format. It is used to get a better grip on the large data sets, which is easy to understand and to make pie charts, histograms, trend-lines, heat-maps, grid-maps and scatter plots. It helps to perform the mathematical transformation, slice and dice data as per the requirement. These analytics capability to analyze data intelligently helps every line of business to make real use of their data. The flexible interface helps to create, share, save and embed the visualized data for fast communication. It provides flexibility to export data and merge with other data sets to quickly discover something previously unknown. Kibana brings the data to life with visuals that can be combined into custom dashboards that helps in sharing insights from the data far and wide. It can then visualize trends and patterns for data that would otherwise be extremely tedious to read and interpret (Elastic.co, 2015).

Elasticsearch and Kibana together makes a very strong time-series analytics platform. Time series horizon is an important variable for this practicum as we have crawled data over the time span of 30 days and based on this data set, we are trying to get the interesting trends and patterns with a user-friendly GUI dashboard. We can feed the data constantly into the stack of Elasticsearch and Kibana over longer period for more accurate results. However, due to the limitation of practicum time allotment, we did our analyzation on the data set of 30 days.

4.4 Implementation

The knowledge discovery process has been described is this section step by step from gathering data to producing the result with respect to our practicum from sources: Daft, Linkedin, Twitter and Google Trends.

4.4.1 Data source 1: Daft.ie

Our major bulk of the data was collected from Daft. It lists various properties for rent and sale available in Ireland. Since the scope of the practicum is restricted to rent in Dublin, we concentrated only in the properties of the capital city Dublin. Also, our analysis would be mainly from the renting point of view. Outcomes of which would be useful for the renters, the real estate agents and likewise. Our main aim is to get as many property details as possible and build correlation among the various attributes of the property. This would help us in drawing valuable insights in relation to the housing market, like the renting behavior of the customer, the duration for which the property would stay in the market, the distribution of the price across the areas in Dublin, etc.

4.4.1.1 Attribute Selection

Initially we studied and analysed the site thoroughly to decide on the keys to be considered for each individual property, which would act as an attribute. After keen observation of the advertisements and discussions with our supervisors and colleagues we decided on the below mentioned terms to be made as the attributes or the parameters for the foundation of the practicum.

We selected the following fields to be considered for our practicum:

- PageUrl: It is an URL of each property called as source page URL.
- Area: Complete address of a property.
- Price: Price of a property with '€' currency symbol in front and the mode of payments i.e. Monthly or Weekly.
- Type: Property type (Apartment/Flat/House/Studio) and has the values of number of Beds and Baths.
- Photos: Number of Images uploaded.
- Views: Number of clicks/ views of a particular property.
- Advertiser: Advertiser complete details.
- Advertised_date: Advertised Date of a property.
- Price History: History of all the changes in price for a property. It is a combination of dates and the price changes on a particular date.

PageUrl would provide the unique id for each property listed on the website. It would be convenient to locate the individual property with respect to their unique ids. **Advertiser** details is needed to help us segregate between private and agent among the advertiser type, it would be interesting to note if people are biased towards the category of the advertisers while renting in. **Advertised date** is the date when the property is put up for advertisement (first listed date). However, Daft has the feature that each day it updates the advertisement date of most of the properties with the current date. Hence, there are few assumptions we have made during our analysis as mentioned earlier. **Type** of property would give details of its type i.e. house, flat, apartment, studio and its corresponding number of beds and baths. For studio the number of bed is always zero. **Views** are basically the number of clicks on the property. This attribute would give the count of the people who viewed the property and is considered to observe if there is any relation between viewing of a property with other attributes such as price, advertiser type, etc.

4.4.1.2 Data collection

Data was collected using specific techniques. We began with our crawling and extracting activity. We also used a tool to enhance the data extraction process named ‘import.io[1]’. We ran the crawler every day from 2nd July to 3rd August given the time duration of the practicum. We made use of a crawler as well as an extractor. The combination of the two was chosen because the data we wanted to extract was split over two pages. The basic data such as area and its price was available on one page but the more detailed data we are interested in like the type of property, number of beds/baths, views, advertiser details were one level deeper. Hence, to retrieve the more valuable data, we made use of a combination of an extractor and a crawler. With the help of the extractor, we first extracted the corresponding urls for every property on the site for rent. For this, we made use of the offset for each Daft page, which had a pre-defined pattern and extracted individual unique ids for each advertisement put up on site. Offset is an integer indicating the displacement from the beginning of the object until a given element or point. It denotes the number of address location added to a base address in order to get to a specific absolute address. The offset for each page was set at a difference of 10. We then trained spider crawler for Daft.ie to get the data that we would want from it. The data we crawled were the attributes mentioned above. We mapped the corresponding data type for each of the attributes while training our crawler.

[1]: https://import.io/

Attributes	Attribute Type
PageUrl	String
Advertiser	String
Area	String
Advertised_date	Date Format (mm/dd/yy)
Price History	String
Photos	Integer
Price	Integer
Type	Integer
Views	Integer

Table 1: Crawled Data attributes and their types

After having trained the crawler, we fed in the urls extracted from the extractor in to ‘where to start’ field of the crawler specification. Accordingly, it will start crawling from the mentioned urls given as the input. In addition, we provided the same set of urls in the ‘where to crawl’ fields so that the crawler crawls through these urls and retrieves the information as per desired format, mentioned during the training provided to them. Also, by doing this we set the parameters of the url pattern of the site we are crawling. It is necessary because once we have the exact urls, the crawler becomes more efficient in retrieving the data it would not have to travel to the unnecessary places to find for the data. Hence, it is quicker in returning the desired output in lesser time. ‘Where to extract’ field requires the url pattern of the site pages. The crawler extracts data from any page that matches the pattern. Also, the url template can be defined for the pages. Since, we have the exact urls from which the data is desired we give these as input in the ‘where to extract’ field as well instead of giving just the pattern of the site. The page depth was kept as zero, as we didn’t want the crawler to access any other links within the sites and this way the output would come faster as fewer the page depth the quicker the data is fetched. In order to restrict the load on the Daft site we put a sleep time of 2 seconds after every page.

Simultaneous pages were set to minimum of three, as we did not want the crawler to crawl through more pages at the same time and increase the load.

Using the urls, we crawled through each and every property listed on Daft. The property ranged from 1800-1900 on a daily basis. We crawled data for each property page wise. The output from the crawler was taken in both CSV and JSON format for further formatting and processing in separate ways. Tableau was used mainly for analysis and Kibana was used for near real-time visualizations and dashboard preparation.

4.4.1.3 Working with Tableau

The CSV file received as an output was further formatted to distinctly segregate the fields of interest. From the addresses of the property, we extracted the last area code (postal code) so that we could classify the areas broadly under the respective area codes. We added crawl date and advertised date which is the day on which the crawling was done and day on which the property was put up on Daft respectively.

We calculated the most important field i.e. the duration of the property in the market, which denotes the life of a property determined by the number of days it is listed on the website before being taken up. We made use of MS Access to make it convenient to extract the required data from the table stored in the database. We used SQL queries to get the data and do some quick calculations. To determine the life for each property that appeared on each day of the crawl until it was taken off the site, we made use of its max crawled date and min advertised date. It implies that for each property, it will look for the max date of the crawl and subtract it from the min advertised date, which is when it first appeared or listed on the site.

The formatted CSV file with the required added fields was used as an input for tableau which helped in visualization and analyzing the trend conveniently. We plotted different charts with various factors and different combinations to analyze the pattern across areas and dates.

4.4.1.4 Working with Elasticsearch and Kibana

Following is the architecture, which we have followed to index the crawled data in JSON format into the stack of Elasticsearch and Kibana.

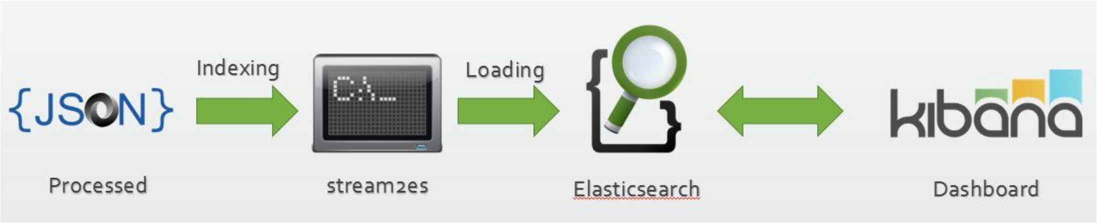


Figure 1: Elasticsearch and Kibana Architecture

The original crawled data was captured in unstructured JSON format as:

Auto-Captured Keys:

- timestamp: day, time and method of crawl
- guid/username/email/roles: Account credentials
- _resultNumber: Numeric orders of crawled properties
- _widgetName: Type of Imports inhouse widget/tool used
- date/_utc: Crawled date

The entire above auto-captured and training model crawled keys (discussed in previous section) needs to be restructured because Elasticsearch can only perform queries and aggregate on the fields with proper data type. For the same, we have first identified the important fields from the crawled data and thought in advance as in what kind of queries and aggregations we would like to support in our model/ dashboard using Kibana. For this structuring of data, we have created a script in python which does the data transformation, data cleaning and data addition such as addition of geo-coordinates for each property in order to make heat-maps and further locations based analyzation. Other data structuring, data cleaning and data transformation job that the python script has implemented is discussed in details in the appendix section 1.

Indexing and Mapping of Data

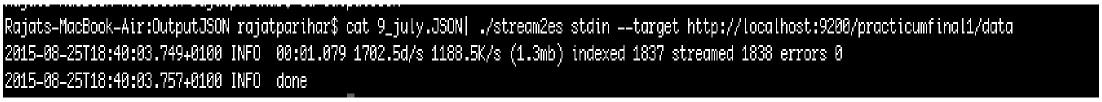
The next step is to create index and to define the mapping of this index as per the data fields of the processed JSON files. Mapping is the process of defining how a document should be mapped to the Search Engine, including its searchable characteristics such as which fields are searchable and if so then how they are tokenized. In Elasticsearch, an index may store documents of different ‘mapping types’ and it allows associating multiple mapping definitions for each mapping type. Here, we are actually telling Elasticsearch how to index each field. The default properties for STRINGS and INTEGERS are accepted by Elasticsearch normally. However, the date fields and the geo-coordinates needs to be declared explicitly to the Elasticsearch along with a method to parse them accordingly. Following are the curl commands to create an index and define its mapping:

```
PUT /practicumfinal?pretty                                      #Creating Index
PUT /practicumfinal/data/_mapping?pretty                        #Mapping the fields
{       "data" : {
              "properties": {                                   #Mapping Date type
                    "AdvertiserDate" : {
                    "format": "dd/MM/yyyy",
                    "type" : "date"
                     },
                     "MaxCrawlDate" : {                          #Mapping Date type
                    "format": "dd/MM/yyyy",
                    "type" : "date"
                     },
                    "Coord": {                                  #Mapping geo-codes
                    "type": "geo_point"
 } } } }
```

Loading Data to the defined Index

To index the processed JSON files, we have used a tool called ‘stream2es’. It is a command line tool, which reads JSON from stdin and it works out of the box. It is a very convenient tool to index thousands of JSON into the Elasticsearch. The following command has been used for indexing:

cat 9_july.JSON | ./stream2es stdin --target http://localhost:9200/practicumfinal1/data



```
Rajats-MacBook-Air:OutputJSON rajatparihar$ cat 9_july.JSON| ./stream2es stdin --target http://localhost:9200/practicumfinal1/data
2015-08-25T18:40:03.749+0100 INFO  00:01.079 1702.5d/s 1188.5K/s (1.3mb) indexed 1837 streamed 1838 errors 0
2015-08-25T18:40:03.757+0100 INFO  done
```

By this way, the processed JSON for each day has been indexed and loaded to the Elasticsearch. In order to analyze the JSON, the indexed patterns were visualized in Kibana.

4.4.1.5 Data Visualization in Kibana

Once the data has been indexed, Kibana was used for visualization of the data. We have identified important plots and made our dashboard with eight visualizations. Three of them are heat-maps, two each of line charts and vertical bar charts and a grid-map. This dashboard has an update interval set for every 15 minutes and it works in real-time on the localhost. It can be customized on the go and is user friendly with the features and filters accessible to all. These access can be defined based on the business requirements. The data is coming from two different sources (Daft and Google Trends) at the backend. Whenever a new data file is processed and loaded into the pre-defined indices into Elasticsearch, then all the visualizations will gets updated on its own in Kibana and this dashboard can be treated as a near real-time live website. The visualizations are created with buyers, sellers and investors perspective, which is discussed in details later in the analysis and discussion chapter.

4.4.1.6 Predictive Model

Predictive modeling is a process of predictive analytics used in the big data domain of business analytics to create a statistical model with forecasting probabilities and better insights. In this practicum, an attempt has been made to identify the patterns and trends hidden within the unstructured crawled data of housing market of Dublin. These patterns are then used to create predictive models that will try to forecast the property prices which will help the stakeholders towards the stochastic nature of the housing market. The classification is used as a predictive modeling technique to make a Decision Tree. The parameters and the data set used are as follows:

- The classification algorithm used is C4.5 (J48 implementation). The experiment is conducted on the Windows 7 x64 OS with 32GB RAM and Quad-Core processor of 4.4Ghz.
- The Predictive Analysis was done on the data-mining tool Weka v3.6 (64 bit edition) and Random Forest has been done on RStudio v 0.99.
- Collection of seven nominal and numerical attributes and the last attribute is the rent price (Price Group) which is used as a class attribute. It has the nominal grouping of prices as < 750€, 750-1500€, 1500-3000€, 3000-5000€ and > 5000€. The seven attributes are life of a property (Life), postal code (Postal Code), property type (Type), advertiser type (Advertiser), number of beds (Bed), number of baths (Bath) and number of views (Views).
- Due to the vast nature of the tree and to accommodate it in the report, the decision tree is prepared on the three main postal codes i.e. Dublin 4 (highest price), Dublin 10 (longest duration) and Dublin 17 (shortest duration).
- The confidence factor is taken as 0.25 and the minimum number of instances per leaf is 2.
- For Random Forest, the seed value is taken as 1234, 90% data is taken as testing data and remaining 10% as test data. 500 trees are generated with 2 variables tried at each split. Complete script is available in an enclosed DVD.
- The accuracy, precision, recall and the other statistical values and the detailed analysis of tree is explained in the Results section.

4.4.2 Data source 2: Property Price Register

4.4.2.1 Purpose

From this site, we have collected the data of all the sold properties for the month of July in Dublin. This was done to assist the Property Owners and Investors to improve their Returns on Investment (ROI) in terms of rent/sell ratio.

4.4.2.2 Method

We have tried to interface the data obtained from the property price register for the month of July with the rental prices for the same month. We calculated the ratio of average selling price for each area for the month of July to average renting price in those areas. This is done with the Property Owners/Investors perspective so that they can make better decisions in buying a property to get the best return on their investments.

4.4.3 Data Source 3: Twitter

4.4.3.1 Purpose

The purpose of collecting data from twitter is to understand users' requirement and their behavior based on their profile and tweets. Commercial point of view is to understand the current trends in housing market by following various real estate twitter accounts.

4.4.3.2 Method

Data was initially collected from 25th May to 10th June from Twitter for about 20 keywords. The purpose of identifying tweets by the relevant keyword was to understand and forecast the user's sentiment on searching a house or on after getting an accommodation. Some of the keywords chosen were 'Dublin Rent', 'House in Dublin', 'Apartment in Dublin', etc. In order to limit the number of tweets and to get more relevant data, the location (Ireland) boundary was implemented in our setup.

The whole setup was initialized on Amazon Web Services and the data was collected 24*7 using Python. This script initialized a secured connection to twitter using two phases of authentication process and collected data in text format.

4.4.3.3 Challenges

Initially the script was running on author’s laptop but then to save the resources and the limitation of the availability of laptop/ desktop, the setup was transformed to the Amazon Web Services (AWS). The advantage of AWS is that anyone can login to it using the credentials provided and modify/edit/rerun script being used. The three months’ student subscription of AWS has been enrolled.

The output has to be continuously monitored because the free use of twitter API has the limitation of the collection of tweets on limited keywords for a time span of maximum two days. But the termination time is uncertain and the continuous monitoring was required.

4.4.3.4 Output

The average tweets collected per day were in the order of 1800-2000 for the predefined keywords. The twitter data for 15 days were analysed and it was concluded that the data is not much of use as it contains various other irrelevant fields such as tweets which are retweeted by the user, replies to a tweet, retweets of a tweet posted by the user, manual replies, etc. In most of the cases, the actual specifications of a property were missing which was expected to be there for better analysis. Unfortunately, much of the results were not useful like we thought it would be. Mainly because there were so many tweets there were not any feasible way to filter for purchase/rent intent of the people. Spammers and self-promoters add to the difficulty by crowding their unwanted tweets in them. For this reason, we ended up spending a lot of time screening through the tweets and scrolling through lots of media sharing posts but could not make any remarkable outcome from it.

4.4.4 Data source 4: LinkedIn

4.4.4.1 Purpose

The purpose of collecting data from LinkedIn was to get a brief idea as in how many job on an average are posted in Dublin across different sectors such as IT, Medical, Finance, etc. Then, the objective was to analyze if the increase in the job opening leads to increase in the rent around that area.

4.4.4.2 Method

The primary objective was to get all job postings for a company based on the company ID. In order to do that, the script has been prepared in python, which makes two separate preliminary calls: one to acquire the company name using the company ID and then a second to acquire all jobs based on the company name. Then another call to the API was made for each job ID returned by the second call.

For the same, the client ID and client secret has been generated and used for the data collection. The authentication method was OAuth 2.0 (LinkedIn in-house security method). The Company API, Search API and Job API were used as a part of our data collection phase.

4.4.4.3 Limitations

The LinkedIn API has a predefined throttle limit and has to be called under this limit. However, the limit has four times relaxation if a developer API accesses it.

The OAuth 2.0 security is hard to implement on batch mode. It detects the automation method and sometimes blocks the application if it tries to access restricted search queries.

4.4.4.4 Output

We successfully managed to get the job postings by a particular company but we could not determine the number of applied applications and their geographical locations. The challenge was to retrieve the geographical location of all the number of jobs that were taken up by international or the locals, which would be helpful in forecasting the demand of the properties. We tried getting this information using the IP address associated with the crawled data. However, due to the data and privacy protection policy of LinkedIn, our script could not capture the user data associated with a particular job opening.

4.4.5 Data source 5: Google Trends

4.4.5.1 Purpose

Google trend has been incorporated into our analysis to see if jobs have any effect on the housing market over the span of our practicum duration. Google Trends provides the searches for a keyword over the time as specified by the user.

4.4.5.2 Method

We have taken data for last 90 days on a keyword ‘**Jobs in Dublin’**. Google Trends is used to determine how often a particular keyword is searched on Google. It has the flexibility to define a time period and the data can be saved in *.CSV format. With Google Trends, we have tried to identify if there is any correlation between the house rental market and the number of jobs in Dublin. The CSV file was converted to JSON and then indexed to the Elasticsearch in a similar manner using ‘stream2es’ command line tool.

Chapter 5: Results

We used Tableau and stack of Elasticsearch and Kibana to visualize the results of our datasets so prepared. Data exploration was done by careful observation of the plots of the attributes and correlating various attributes together.

Initial plots were made in Tableau to fetch the preliminary information of the housing market during the duration of our practicum. It would sight the trend currently prevailing in Dublin Housing Sector with regards to the price and availability, type and then we further drill down into areas of interest for detailed analysis in the analysis and discussion chapter 6.

Also, the report in the form of a real-time dashboard has been presented in Kibana by carefully selecting the important charts and plots which would cater to the needs of the stakeholders so defined.

Selected plots have been displayed along with their interpretation.

5.1 Observations

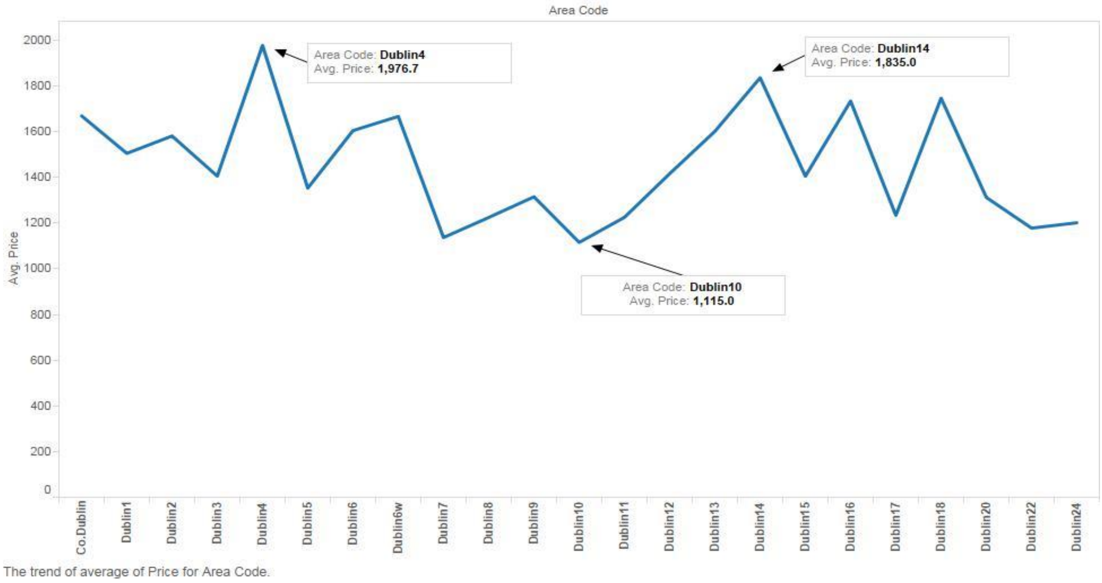


Figure 2: Average price of the properties area wise

The Figure 2 shows that Dublin 4 has the highest average price of €1976 monthly followed by Dublin 14, 16, 18, the lowest prices of the properties are in Dublin 10 with an average of €1,115 monthly followed by Dublin 7 and 22.

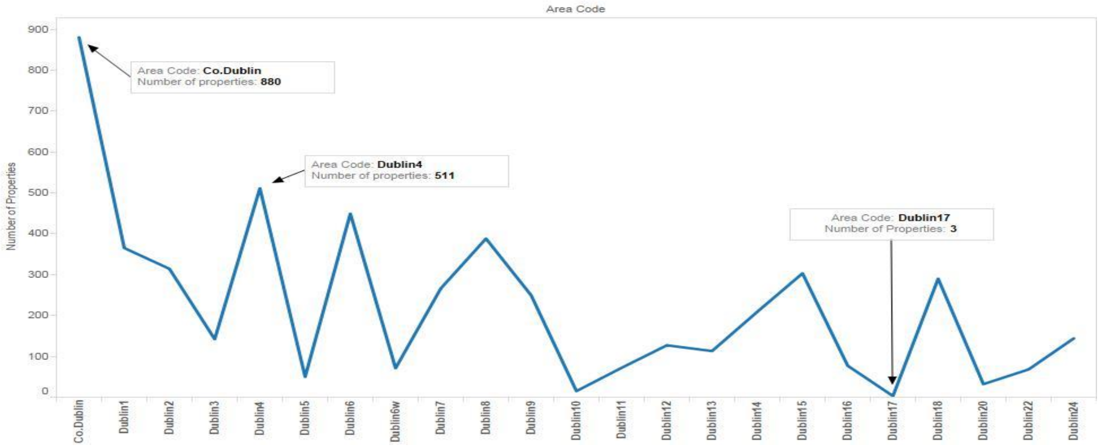


Figure 3: Number of properties area wise

As per Figure 3, the maximum number of properties available as per Draft's listing are in Co. Dublin with 880 properties. Followed by Dublin 4, 6 and 8 in the range of 400-500. Dublin 15 and Dublin 18 also has significant number of properties. Least available properties are in Dublin 10, 17, 5, 6w that is less than 100. Another important data to note was the number of incoming property day wise. We could see that more properties are listed around the mid-week of the month with more or less constant supply throughout the month. In a month there are approximately around 4500-5000 listings for rental section in Dublin.

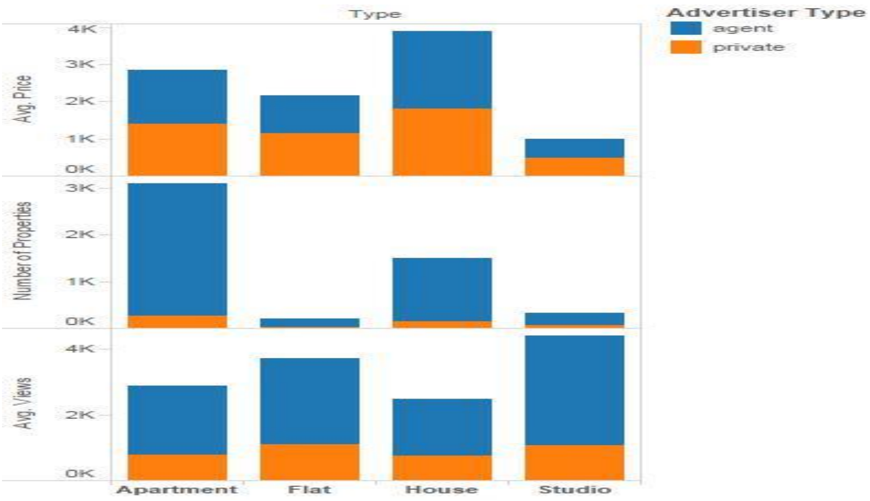


Figure 4: Details of Property Types based on Advertiser type

As per Figure 4**,** the number of apartments available are highest overall, followed by houses, flats and studios. Houses are the most highly priced then comes apartments and then flat and studio. Agents' holds more number of properties than the private advertisers and properties of the agents are priced slightly higher compared to that of the private advertisers except for flats.

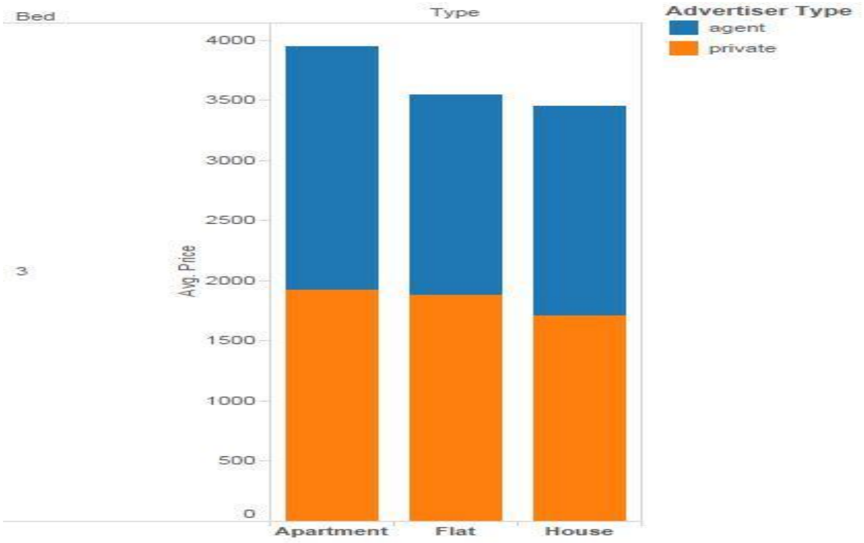


Figure 5: Average Price of 3 bed property based on Advertiser type

For instance for a three bed apartment as shown in Figure 5, the agent charges an average of €2024 monthly and private charges around €1922 monthly on an average. For three bed flat the monthly average price by an agent is €1660 and that of private is around €1883. For three bed house agent charges around 2% higher than private. For studio, agent's rate is around 7% higher than those of private advertisers. Also, the views i.e. the number of clicks on the advertisements are higher for the ones advertised by the agents than by the private.

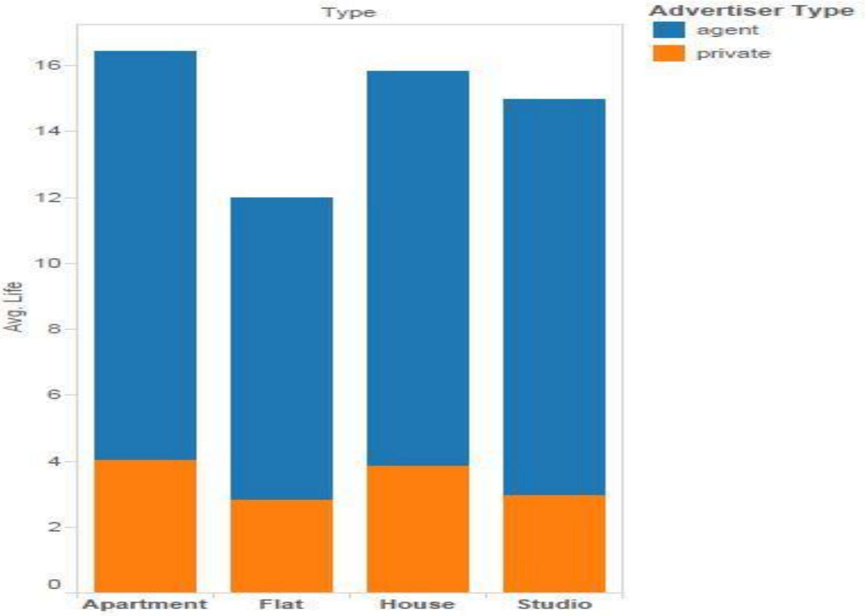


Figure 6: Average Life of the property types based on Advertiser Type

As shown in Figure 6, it is also observed that the properties listed by the private advertisers are taken up faster compared to the same posted by the agents.

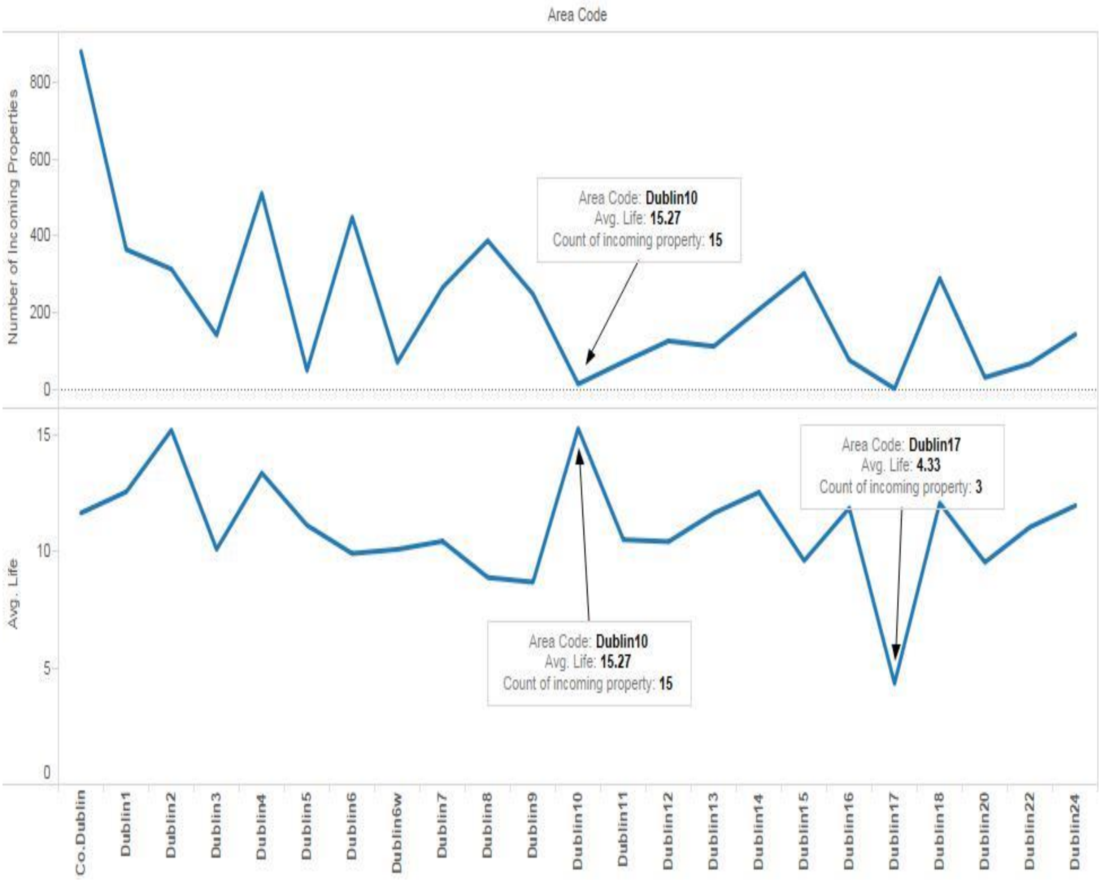


Figure 7: Number of Properties & Average Life Area Wise

As per Figure 7**,** properties in Dublin 10 stays for the highest duration in the market even the incoming property is the least in Dublin 10 comparatively. It can be inferred Dublin 10 is not under high demand areas. Dublin 17 (Near Airport) has the shortest life of a property. However, it has very few incoming properties. More properties can be put up in this area as the rate of outgoing is promising at the moment.

5.1.1 Types of apartment, their average price and their availability area wise

This was plotted to observe if the availability of the property has any influence on the prices. If a property type is less in number then is the price higher in that area compared to others where the availability is more. Also, it could be just an indication as to which type of property is not preferred in which areas based on its availability and price. However, we found the count and the price of the properties are mostly proportional for most of the areas. Exceptions are mainly observed in Co. Dublin which is discussed in details as follows:

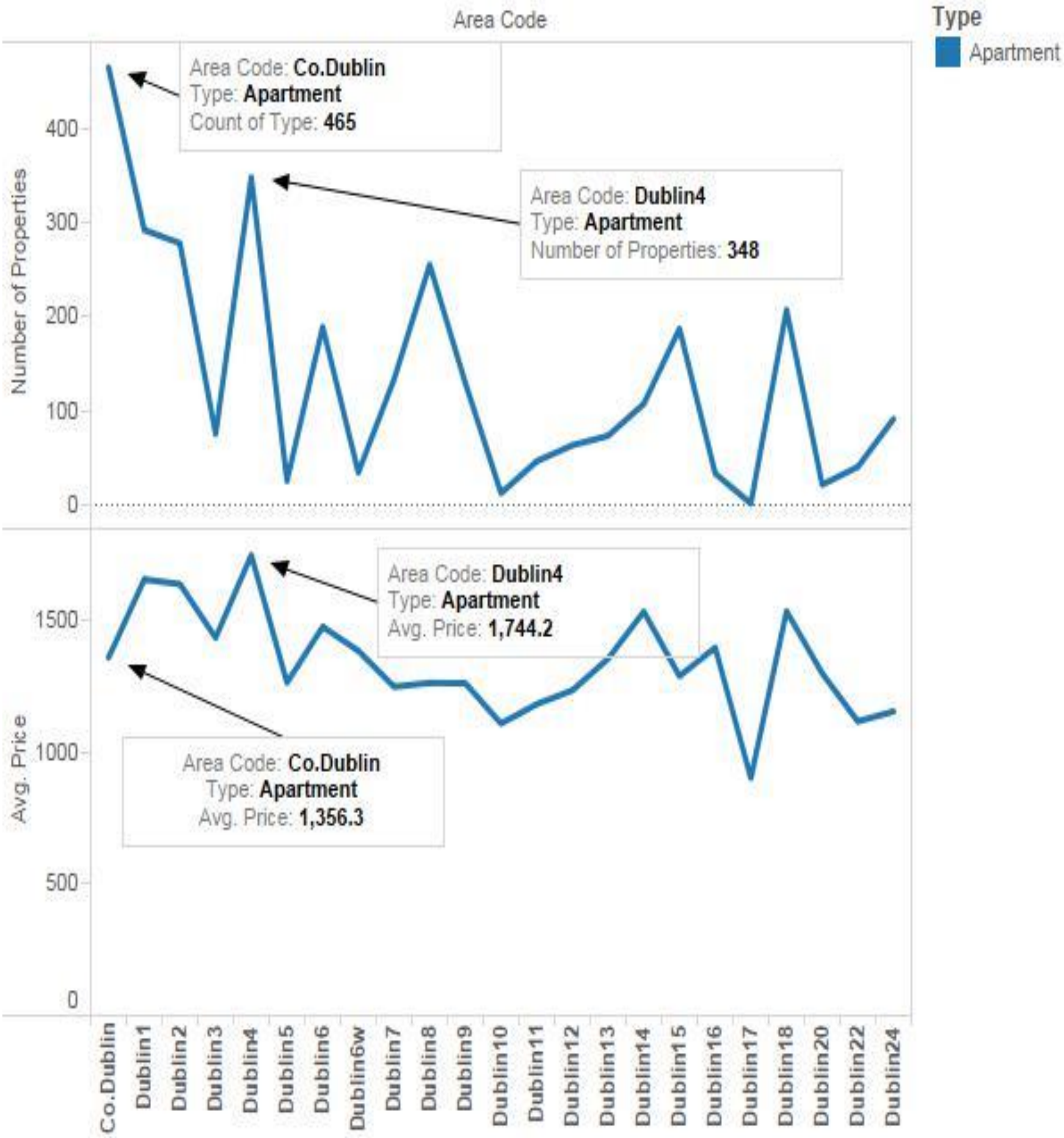


Figure 8: Number of Apartments & Average Price Area wise

Apartment:

Figure 8 shows that Co. Dublin has the highest number of apartments and also the price is low compared to other areas. However, the trend is not very evident in all the areas especially Dublin 4, which has the second highest number of apartment, also has them priced the highest at €1744 monthly followed by Dublin 1 and 2 with difference of around €100. Dublin 10, 17 has the least number of apartments and priced among the lowest. This could be an indication that apartments are not so preferred in these areas.

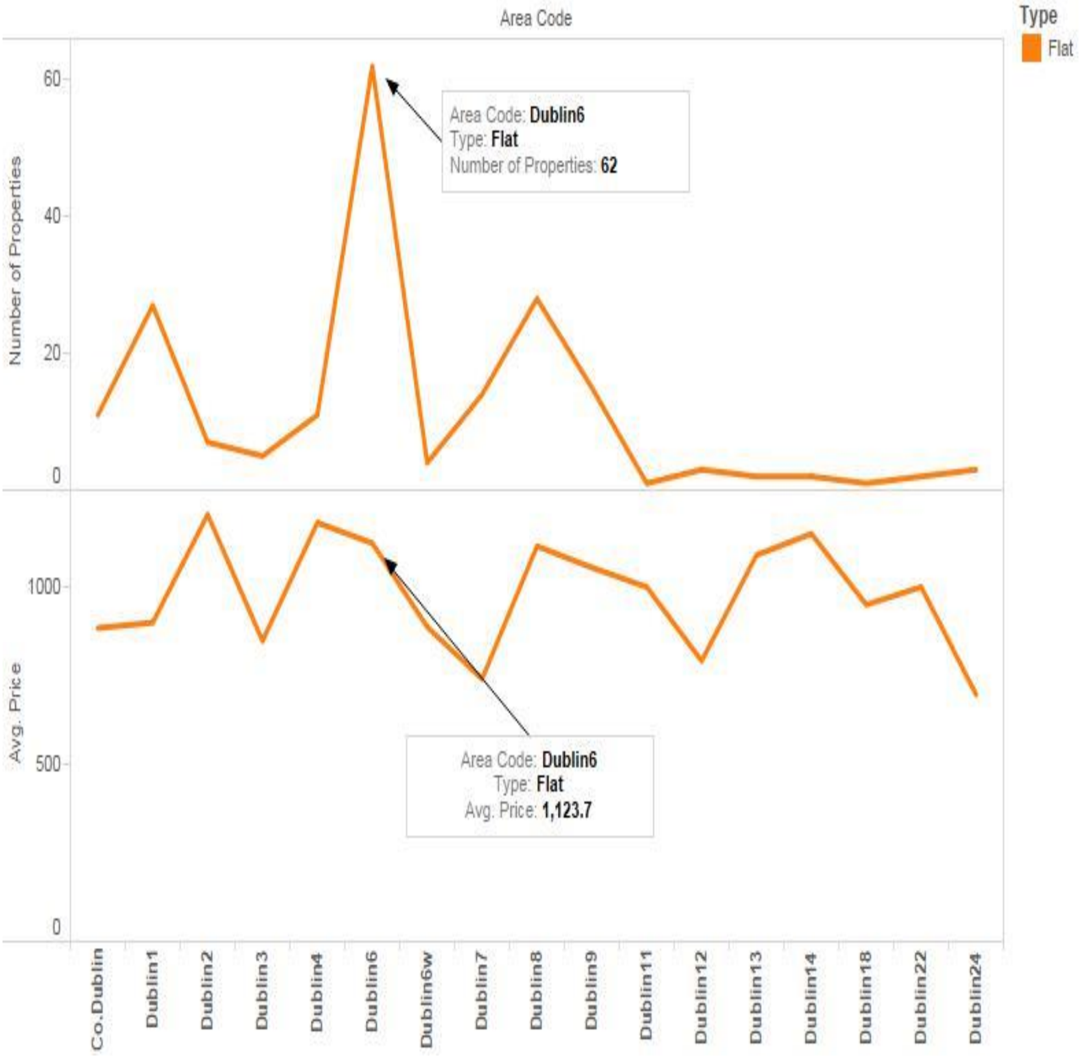


Figure 9: Number of Flats & Average Price Area wise

Flat:

Figure 9 shows that the highest count of flat is in Dublin 6 priced the third highest; Average price is the highest in Dublin 2 followed by Dublin 4, 6 the least price being in Dublin 7. Flat doesn’t seem to be much in demand in areas such as Dublin 7, 11, 14, 24 solely based on the meagre count and low price.

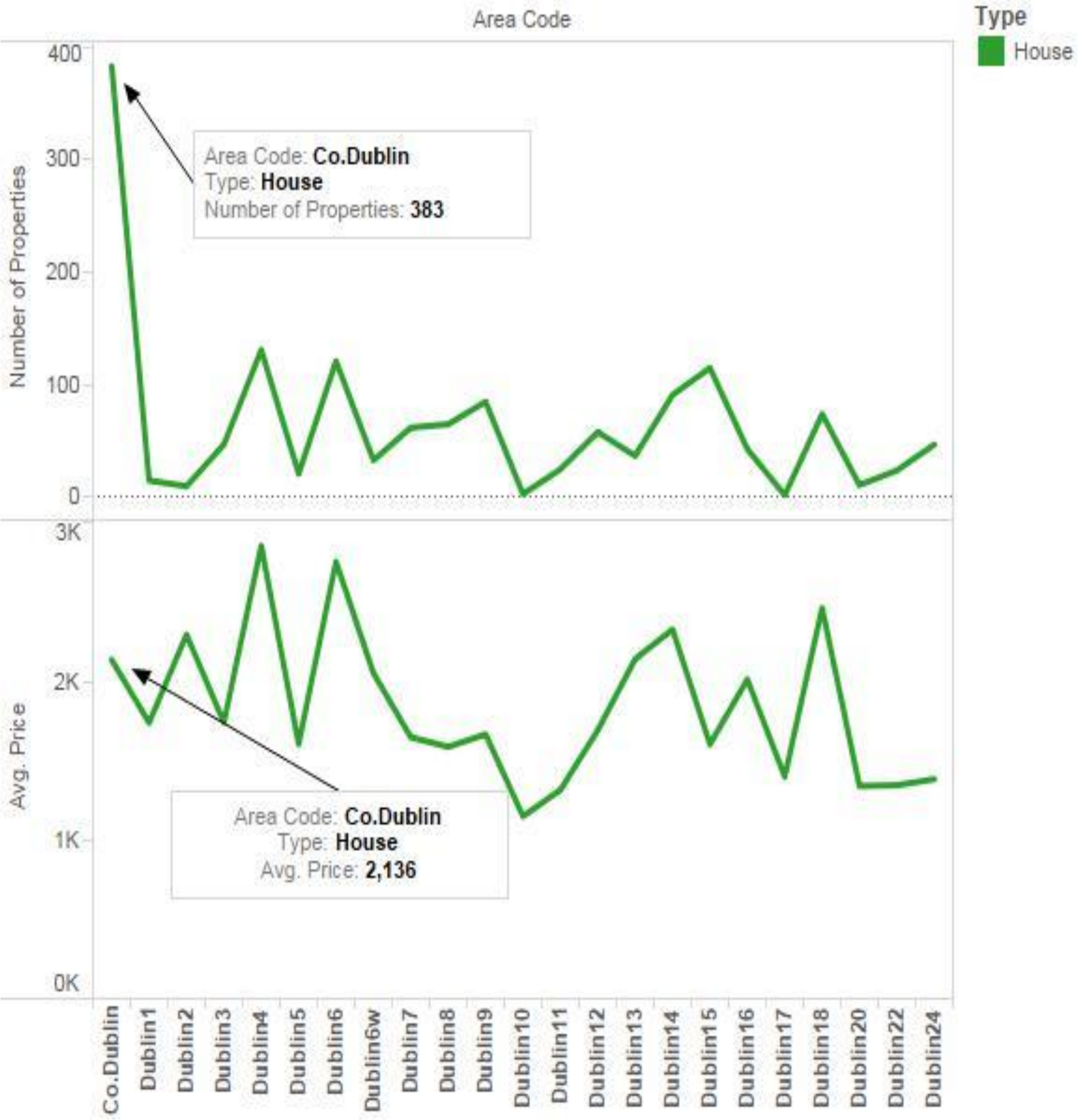


Figure 10: Number of House & Average Price Area wise

House:

Figure 10 shows that the maximum number of houses are in Co. Dublin with average price of €2136. Price is the lowest in Dublin 10 also has the lowest number of properties in Dublin 10. It has been observed that the number of properties and their average price are proportional to each other except for Co. Dublin.

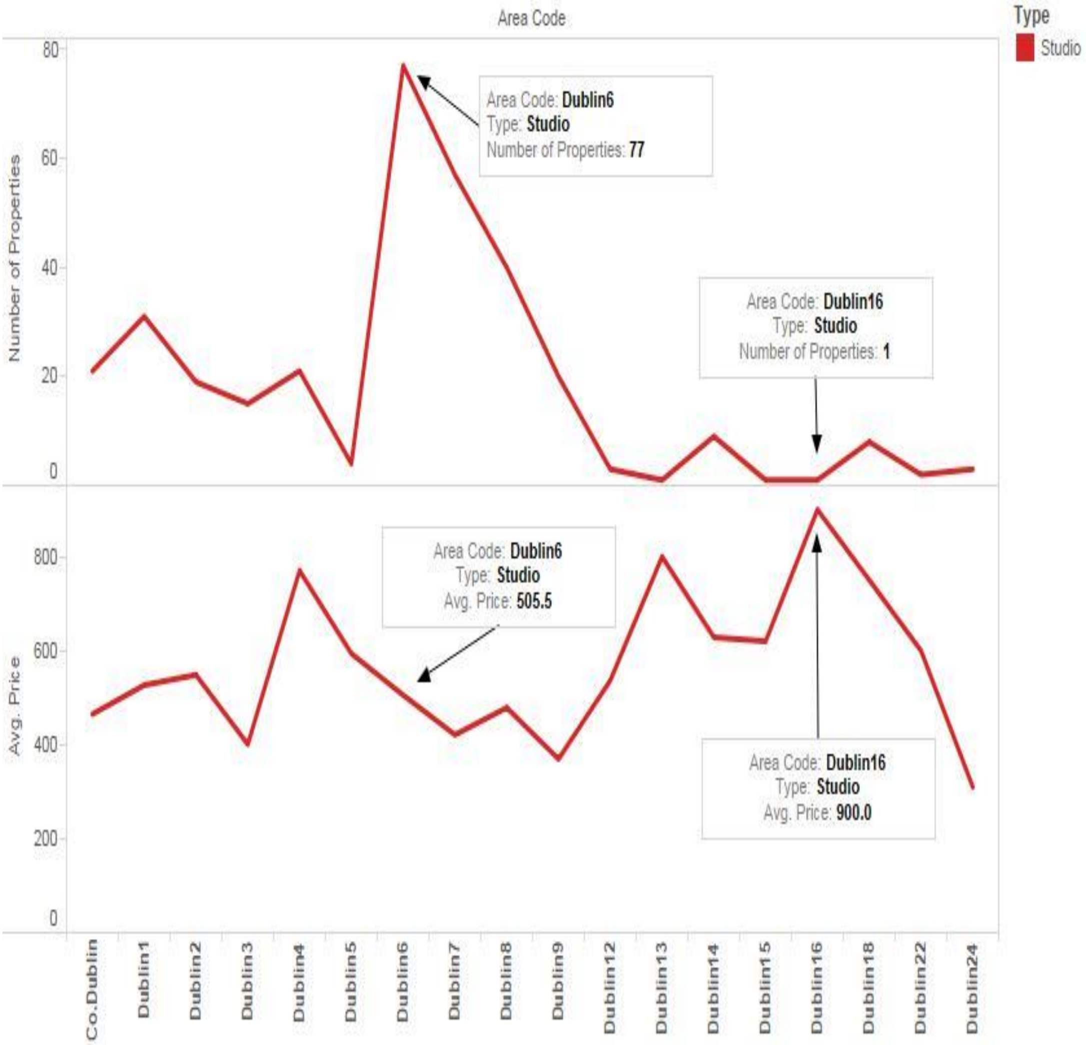


Figure 11: Number of Studio & Average Price Area wise

Studio:

In case of studios as shown in Figure 11, the trend is visible that more the number of flats available lesser is the price and vice versa. Maximum number of studios are in Dublin 6, price being among the lowest of 505€. Most expensive flats are at Dublin 16 the count available being the least.

5.1.2 Number of Property Types based on the number of beds

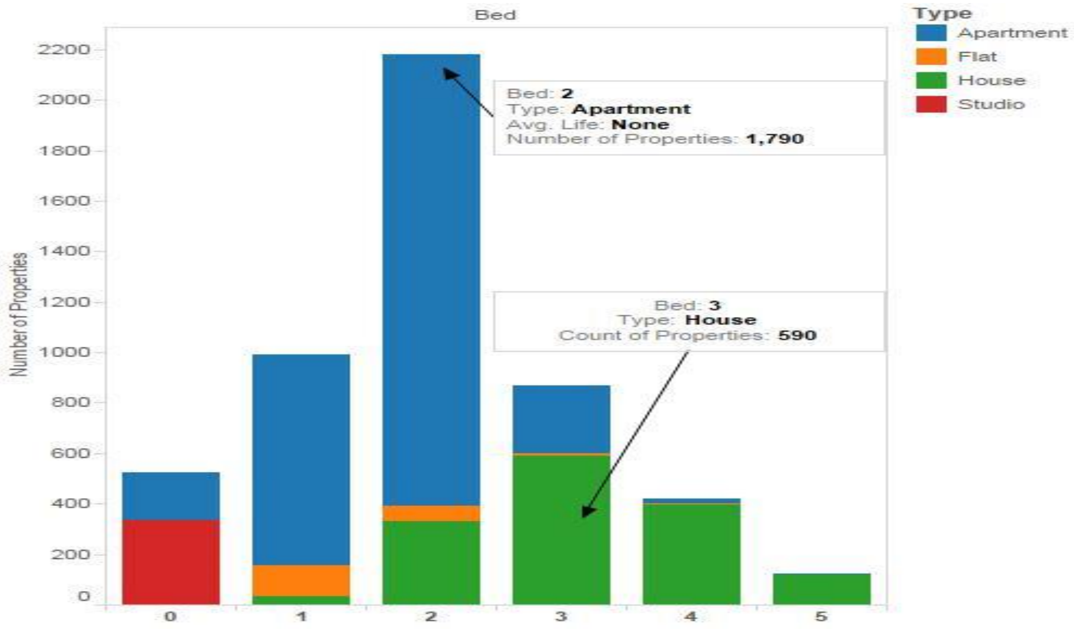


Figure 12: Number of Property Types based on the number of Beds

Figure 12 shows that the two bed apartments are more in number. Followed by three bed houses. There are very few properties available with more than five beds. Studios are higher in number compared to the total number of flats.

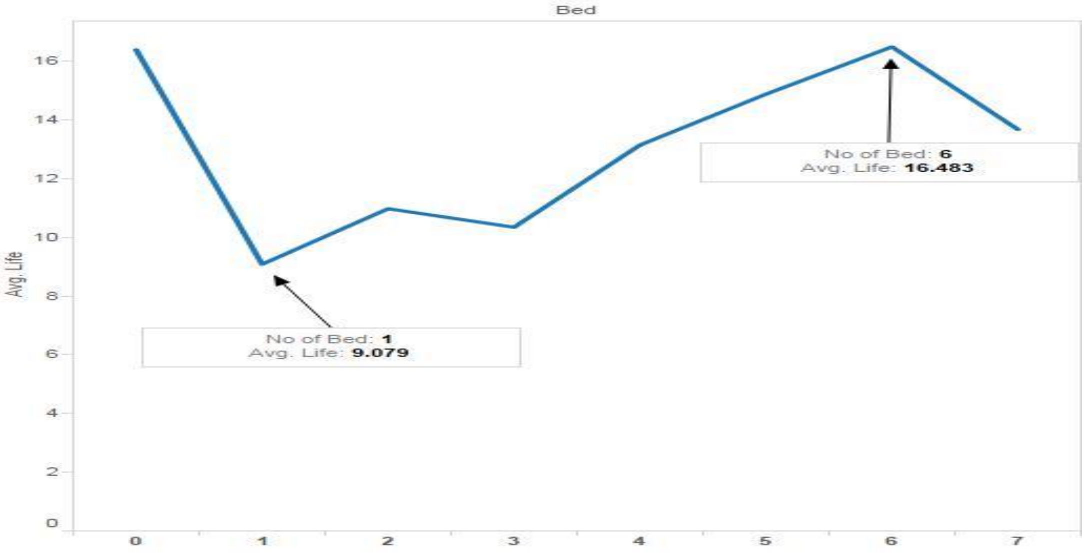


Figure 13: Average life of the number of Beds

Figure 13 shows that the one bed are taken up the fastest from the market followed by two bed and three bed. The slowest movement is in zero bed (Studios) and six bed property with average life of about 16 days.

Movement of the properties area wise based on its type

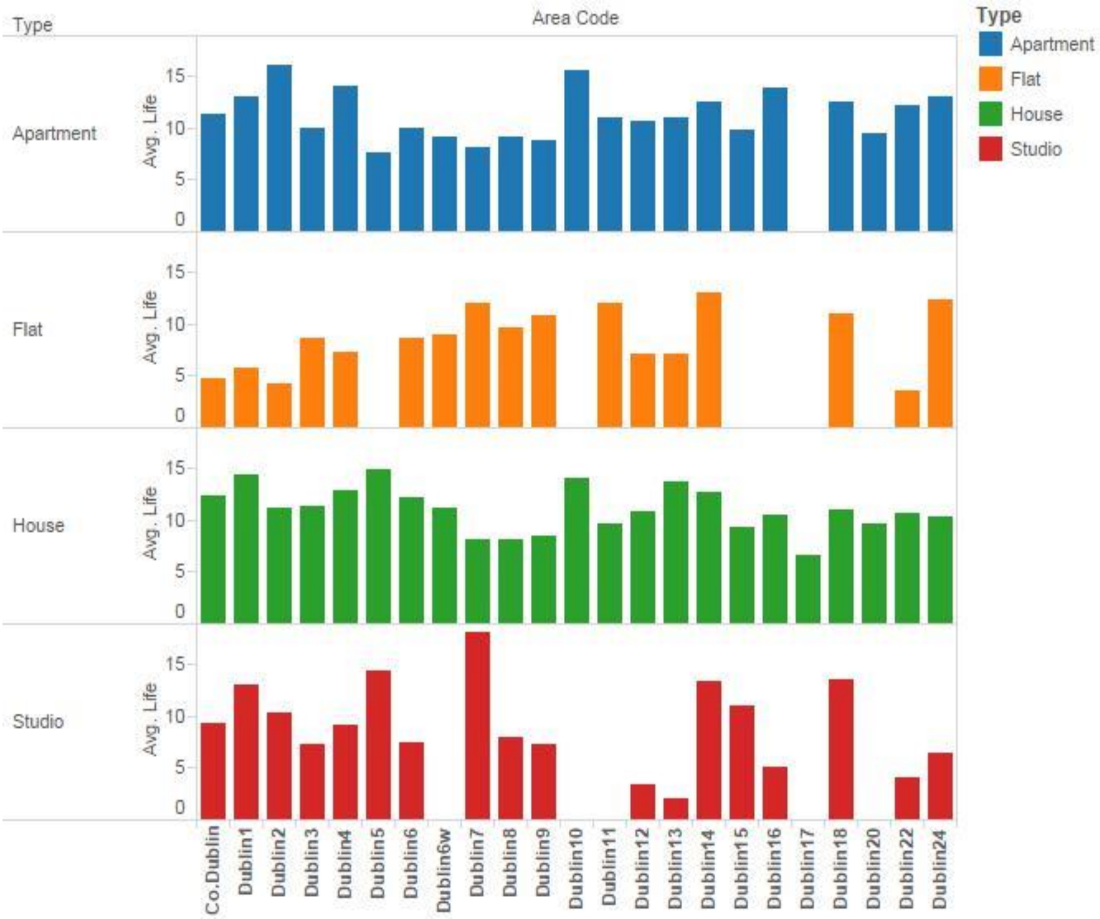


Figure 14: Movement of the Properties based on Type Area Wise

Figure 14 shows the movement of properties based on the property type across all the areas. In case of Apartment, Dublin 5 are taken up faster with an average duration of 7.5 days followed by Dublin 7 and 9. The movement of the apartment is slowest in Dublin 2 and 10 with average life of about 15 days.

In case of flats, the life is the shortest in Dublin 22 and 2.People prefer flats in these areas compared to other types of properties. The slowest movement of flat is in Dublin 14 with average life of 13 days.

In case of house, Dublin 17 has the highest movement for Houses. Slowest being in Dublin 5. Studios have the shortest life in Dublin 13 (2 days) and longest in Dublin 7 (17 days).

5.1.3 Choice of Area based on Type and Number of Beds

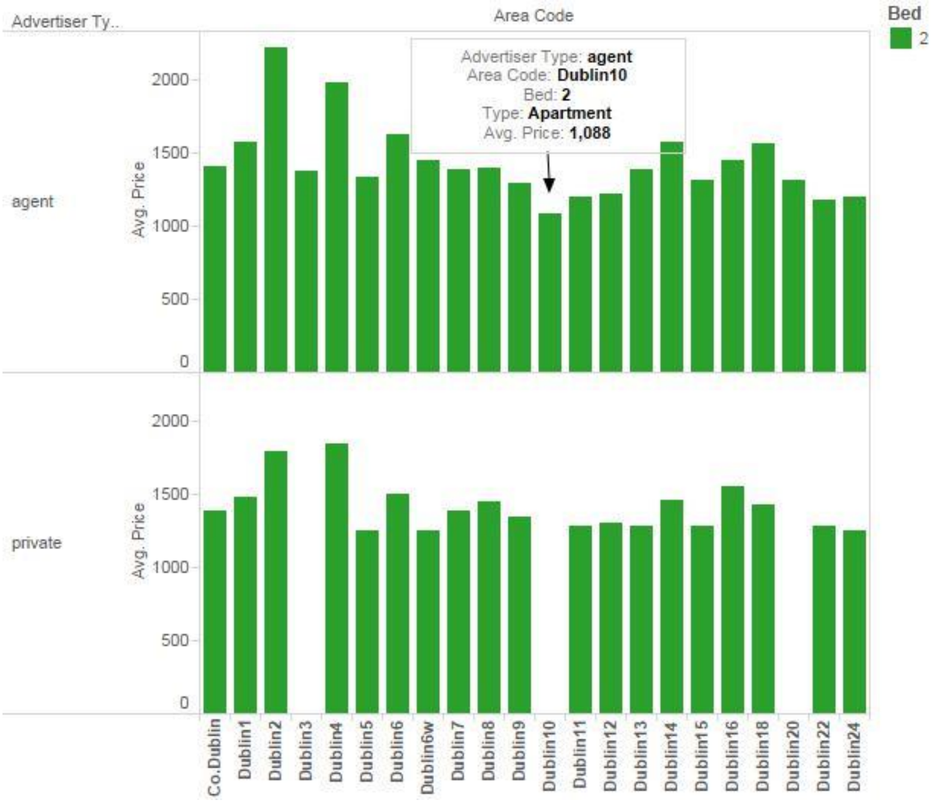


Figure 15: Average price Area wise based on Advertiser Type

As per Figure 15, if the renter has the type of property along with the number of beds in mind then with the help of filters provided he can select the combination of his choice and look across all the areas and select for the area with cheaper price and can decide whether to opt for a property from agent or private by comparing their prices across areas. For instance for a two bed apartment one should preferably look for properties in Dublin 10 offered by an agent which is the cheapest provided there are no other constraints but only the price.

5.1.4 Views vs Area

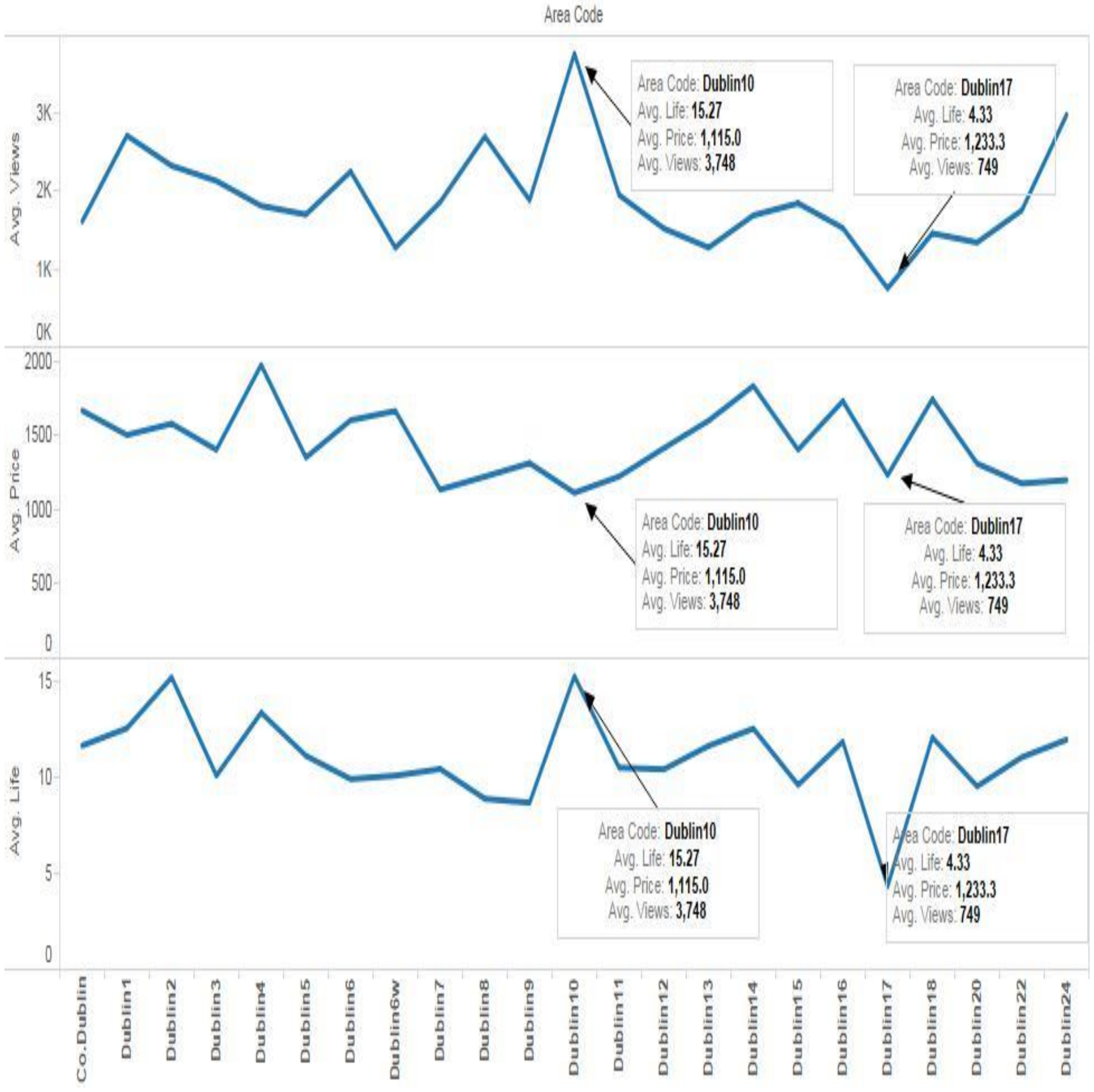


Figure 16: Average Views, Average Price and Average Life Area Wise

In Figure 16, views give an indication of the interest of the people. Views are more in Dublin 10. On an average, more people look for properties in Dublin 10 but still surprisingly the average life there is the highest. Though, people are looking for properties there but they aren’t being taken up by them. We would analyze the case in detail in the later sections. Views are lowest in Dublin 17, still properties move out faster as the area is comparatively cheaper. Views are not directly proportional to the life of the property. For many properties, even if the view is more the life remains high. Either the options available for the property they have is really less or the price is comparatively higher.

5.1.5 Supply vs Demand

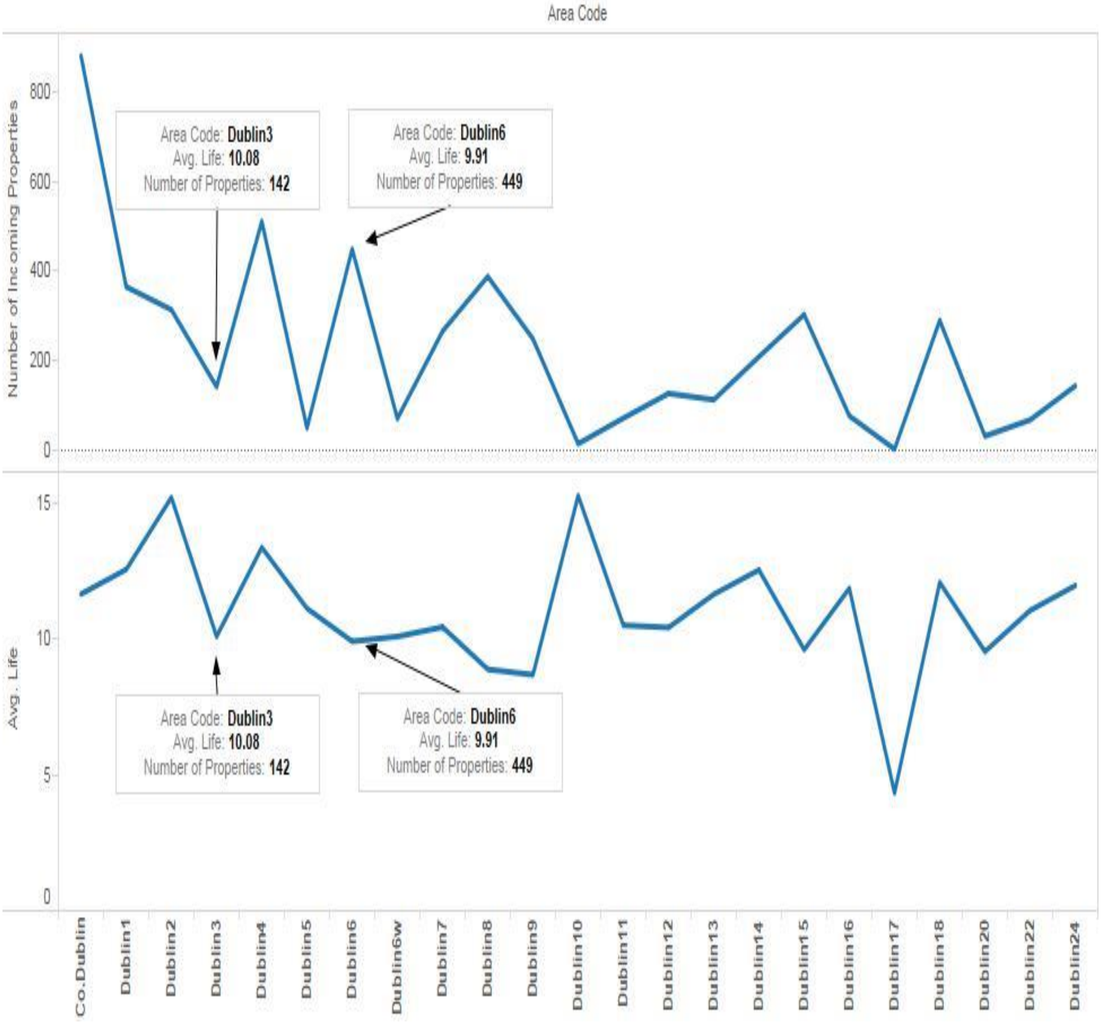


Figure 17: Incoming Properties and Average Life Area Wise

Figure 17 shows the supply and demand plot. We estimated supply by taking the number of incoming properties per area over the month and estimated the demand by plotting the average life of the properties area wise. When compared among all the properties in each of the area over the month if the life of a property is less means the area is in higher demand and if the incoming property in that area is less that indicates it suffers from the shortage in supply. For instance, In Dublin 3, the average life is 10, and the supply is 142 for the same average life of 10 days in Dublin 6 the count of incoming property is 449, which shows it has ample supply whereas Dublin 3 has inappropriate supply to demand ratio and suffers shortage in supply. Similar, problems of shortage are being faced in Dublin 5, 6w and 12.

5.2 Kibana Dashboard

The near real-time report that is available in the form of a dashboard is explained in details in this section. The Kibana dashboard shows the basic and the most important information related to housing market of Dublin. As discussed earlier, it updates in real-time whenever a new data is indexed into the Elasticsearch under the same index name. It has total of 8 visualizations as shown in Figure 18 which shows the following insights:

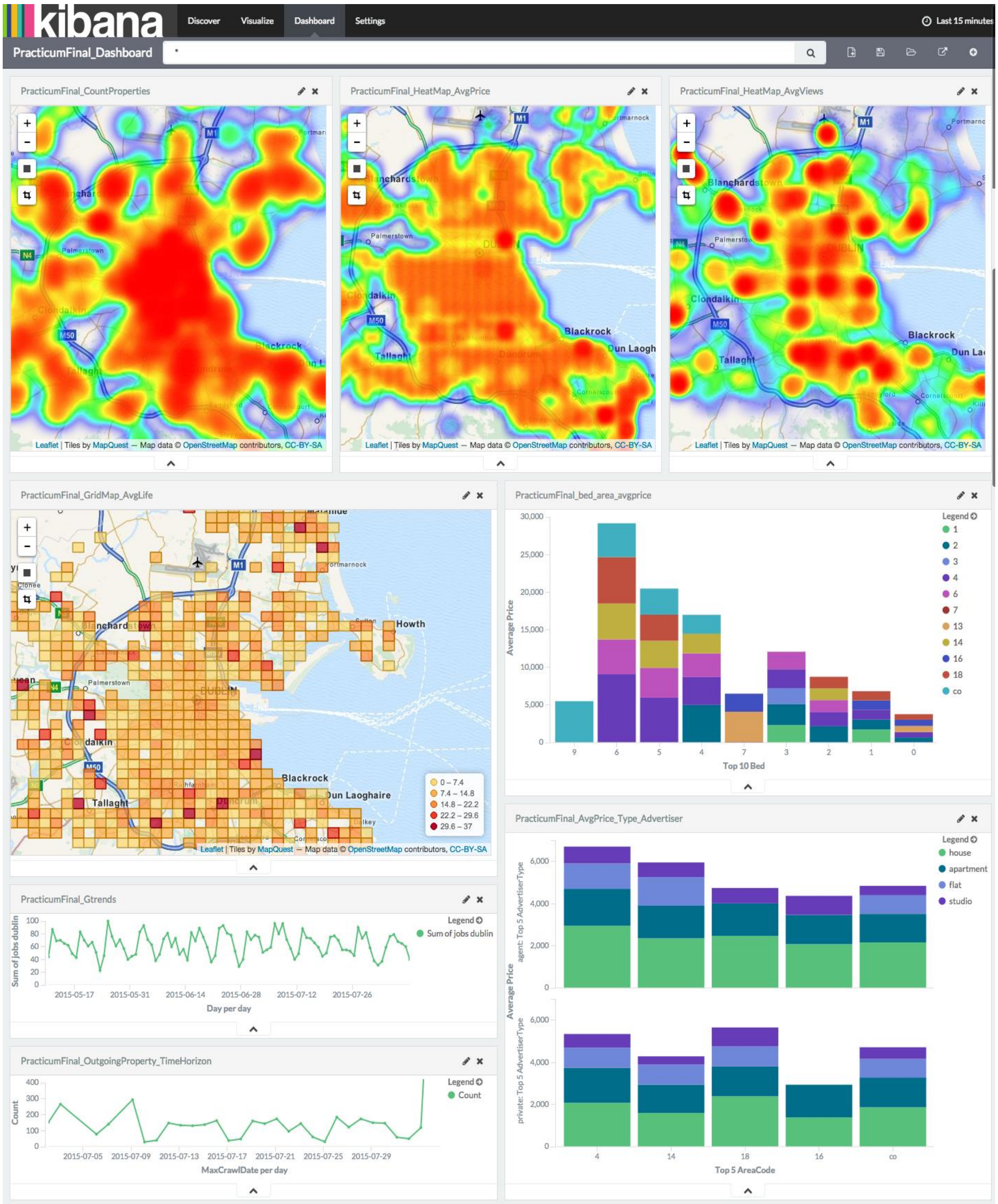


Figure 18: Real-Time Dashboard with 8 visualizations

PracticumFinal_CountProperties

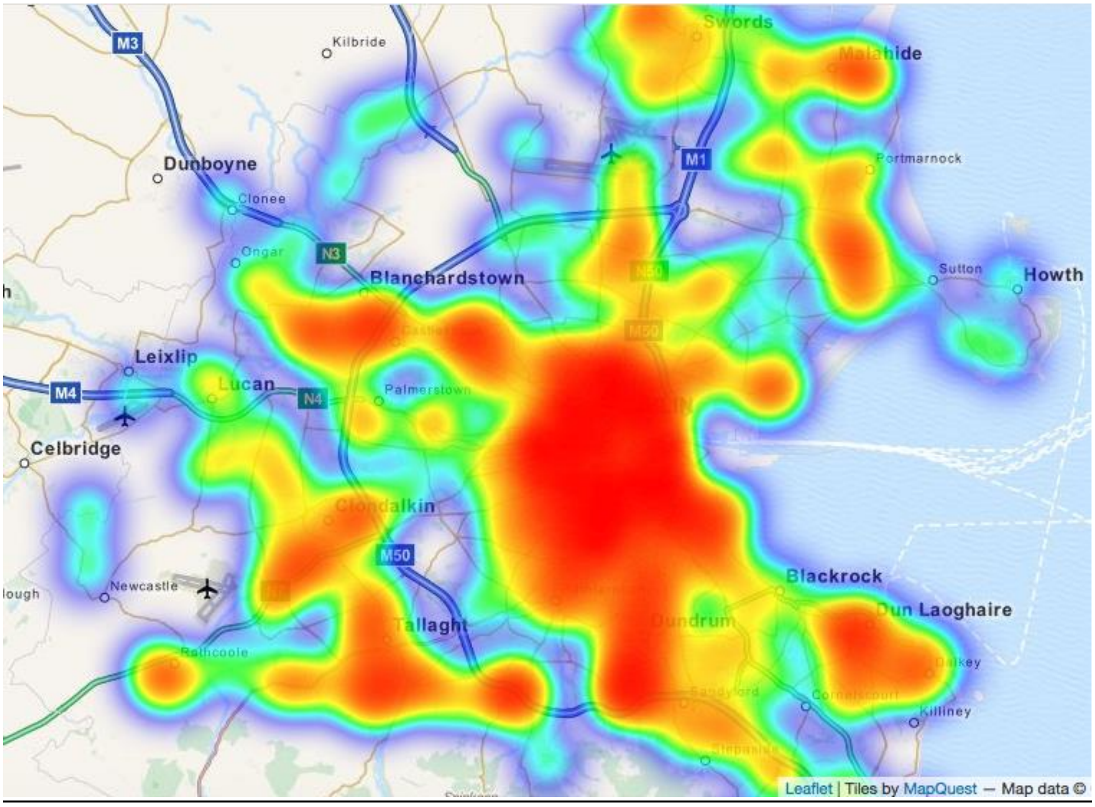


Figure 19: Heat-Map of Number of Properties

It shows the heat-map of the count of number of properties listed in Dublin. This count is the total number of the properties listed during the month of July-Aug 2015. The maximum number of properties available to rent are mostly in the central Dublin, south of central Dublin, Dublin 3 in northern side and some parts of Dublin 15. The number of properties are also high in the North West areas such as Blanchardstown and Castleknock, etc. Other than that, South Co Dublin areas such as Blackrock, Seapoint, Monkstown and Dun Laoghaire also fall in the high availability areas in rental property list. The West and South West Dublin has the least number of properties followed by some areas in North and South of Dublin. As mentioned earlier that this dashboard works in real-time, which means that whenever we load the new data, these plots will update on its own.

PracticumFinal_HeatMap_AvgPrice

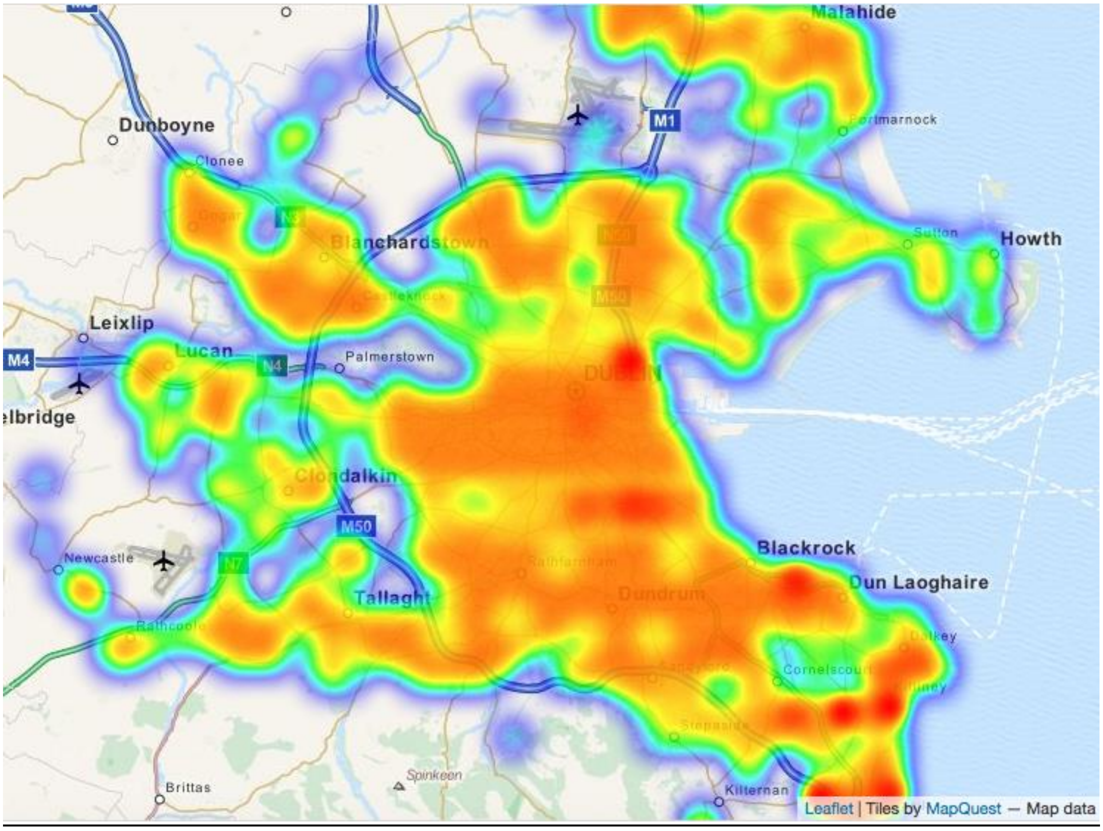


Figure 20: Heat-Map of Average Price of Properties

It shows the heat-map of the average price of these properties. The average price is highest in the south county Dublin i.e. area of Blackrock, Dun Laoghaire and Dalkey followed by south side of city center (near Trinity college), Dublin 6/6w (near UCD college), Dublin 4 (Sandymount, Ballsbridge, etc.) and IFSC, Docklands. We have most of the major IT companies such as Google, Accenture, etc. are situated near IFSC and Docklands. The reason for the average price being highest in south county Dublin is because most of the properties in these areas are big in size and has high number of rooms compared to other areas.

PracticumFinal_HeatMap_AvgViews

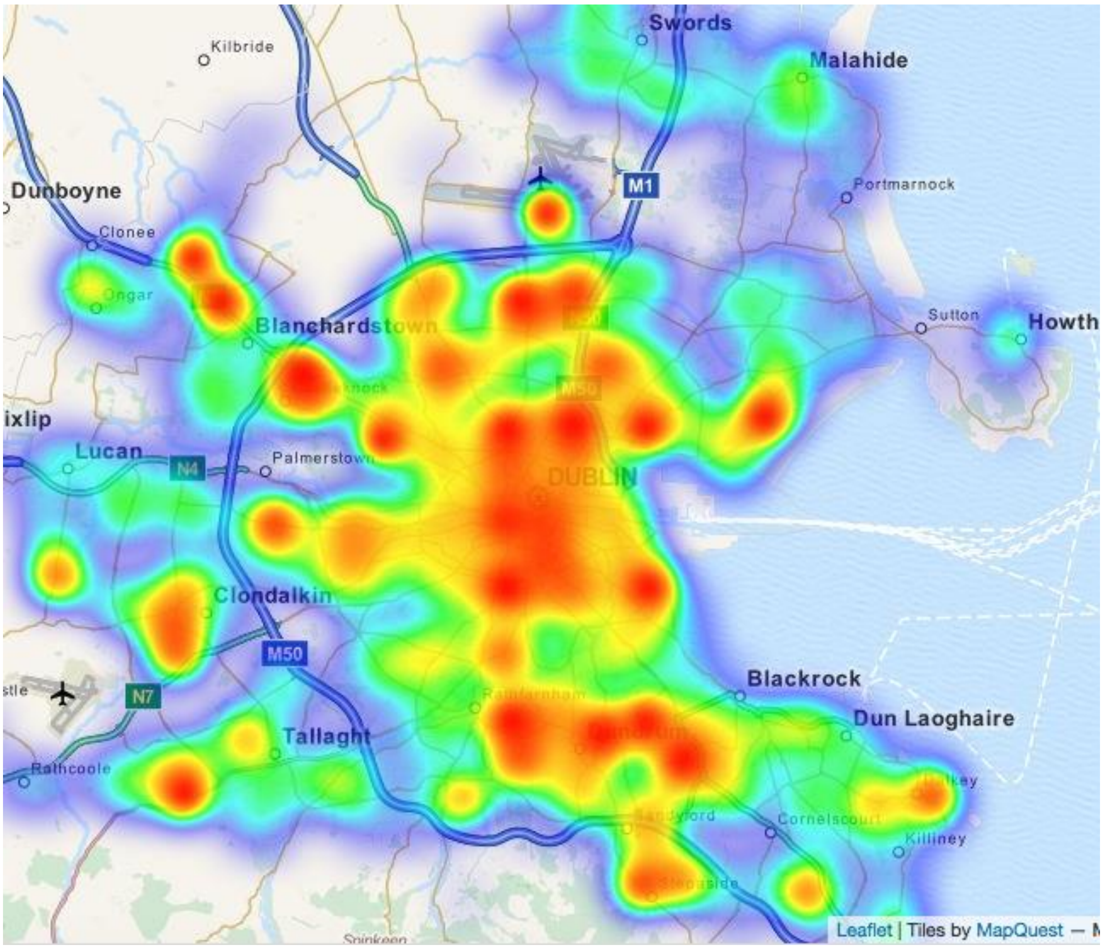


Figure 21: Heat-Map of Average Views

It shows the heat-map of the average number of views of these properties. Most of people have seen the properties where the average prices are less. In addition to this, views are high for the properties near to the three colleges (Trinity, UCD and DCU), North and South City Centre, Dublin 3 (Clontarf), Dublin Airport, South Co Dublin (Stillorgan, Sandyford and Dalkey) and Dublin 8 on the west side. Dun Laoghaire and other South Co Dublin areas are not viewed extensively because of their high price. Very few people have viewed properties in Malahide (North Dublin) even though it is decently priced.

PracticumFinal_GridMap_AvgLife

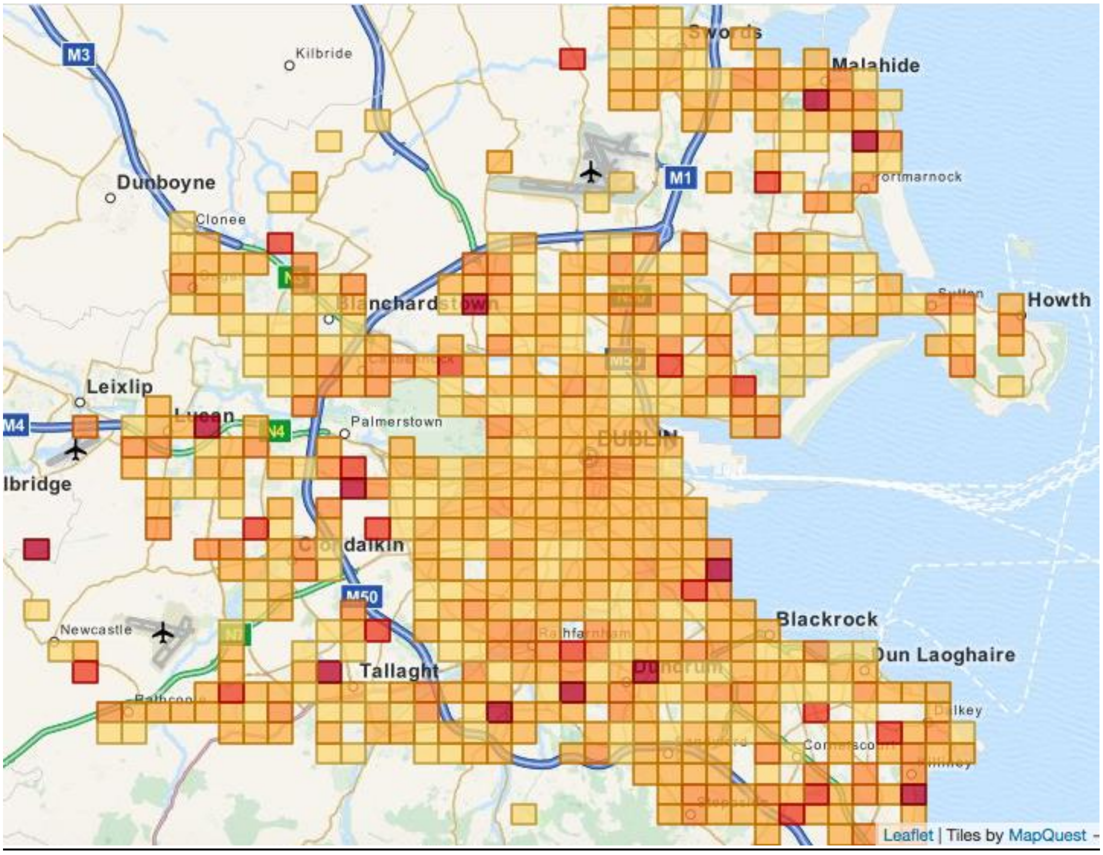


Figure 22: Grid Map of Average Life

It shows the grid map of the average life of properties in days, which is the total number of days a particular property was listed in market before being taken up. The darker grid represents maximum life (29-37 days) and the lighter grid stands for minimum life (0-7 days). Legend is available in main dashboard window. We can see that the properties near the City Centre, Dublin Airport, Dublin 3, Sandyford, Blanchardstown and South Co Dublin such as Blackrock and Dun Laoghaire are going at a faster rate (in less than 7 days) compared to the West and North of Dublin and some areas of Dundrum where average life is about 35 days.

PracticumFinal_Bed_Area_AvgPrice

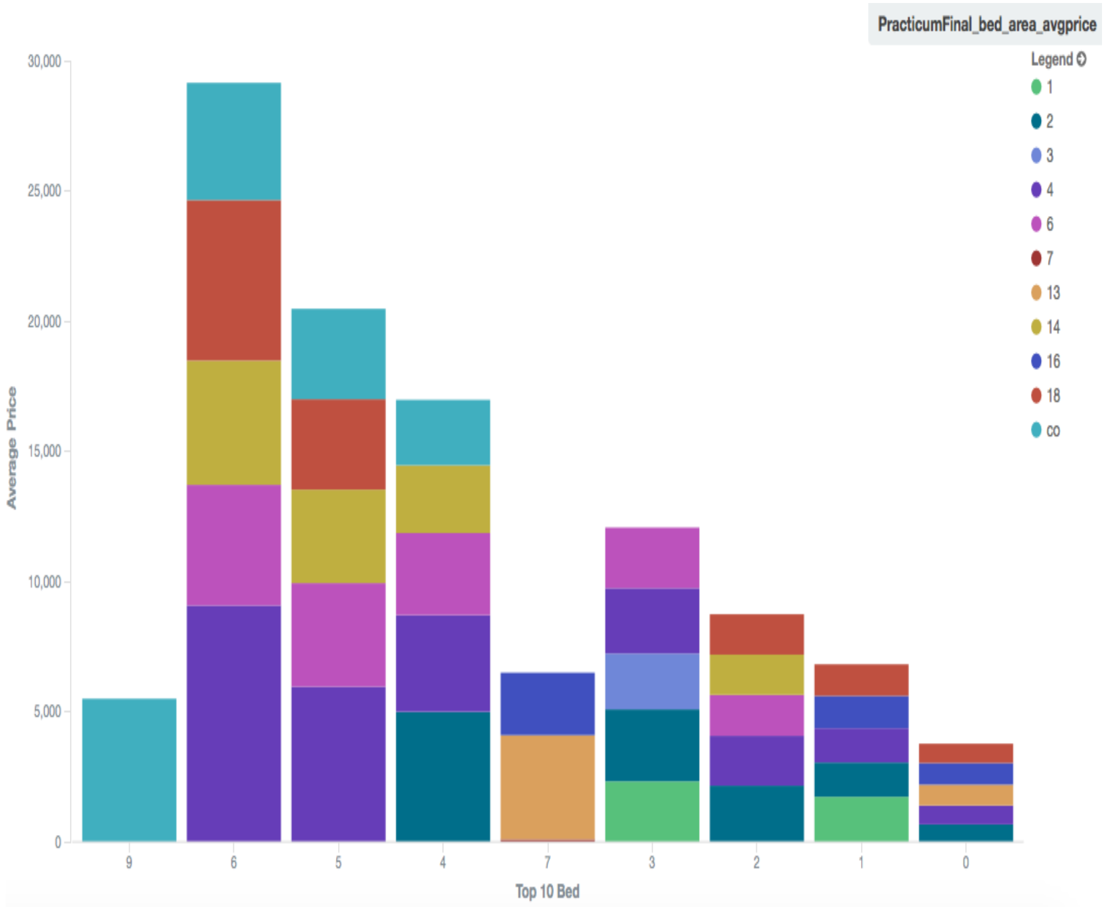


Figure 23: Average Price of number of Beds Area wise

It is an interesting plot as it shows the three different insights on a single chart. The y-axis of this vertical bar chart shows the average price of a property in terms of the number of beds on x-axis across the top five area codes displayed by different colors. We can see that the 9-bed property is only available in Co Dublin with an average price of 5,500€. The average price of two, five and six beds are highest in Dublin 4 and it has also the second highest average price in four beds segment after Dublin 2. The average price of studios (zero beds) are almost similar across the top five postal codes. With this plot, one can decide the area to rent based on the preferred number of beds and their corresponding price.

PracticumFinal_AvgPrice_Type_Advertiser

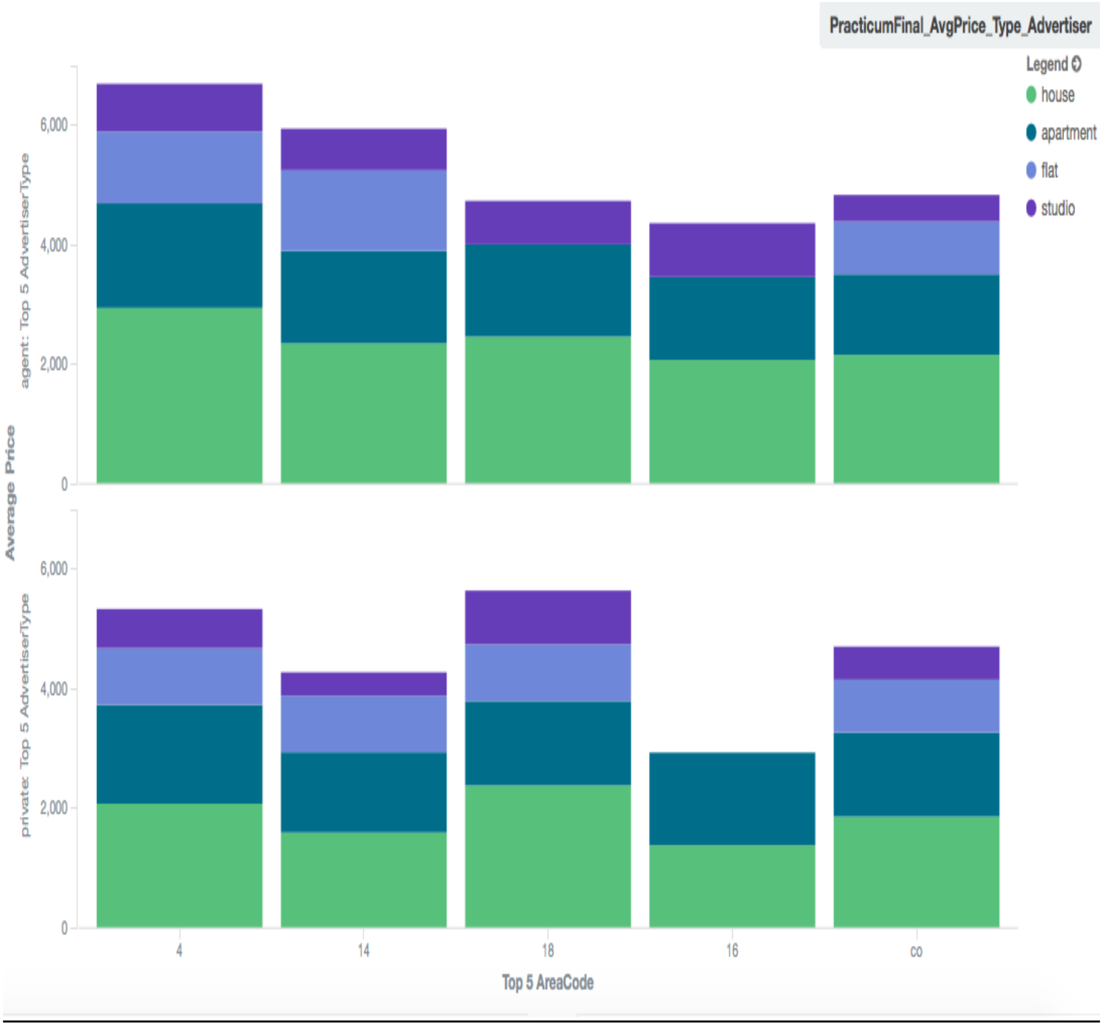


Figure 24: Average Price of different properties type by Advertiser Area wise

It shows the vertical bar chart of the comparison of the average price of different property types (Apartment/ Studio/ House/ Flat) posted by an agent and a private owner. The average price of a house in Dublin 4 and 14 posted by a private owner are about 33% cheaper than the same posted by an agent in the same location. Likewise, the apartments posted in both of these locations by a private advertiser is about 7-10% cheaper compared to the same posted by the agents. It also shows that the private advertisers have not posted any studios on rent in Dublin 16. The trend shows that overall the average price of a property is always cheaper by the private advertiser inventory compared to the properties posted by the real estate agents.

PracticumFinal_GTrends

It shows the line chart of the Google Trends for the number of searches for the keyword ‘Jobs in Dublin’ over the last 90 days. The result from Google trend has been displayed in a tabular format as well for clear understanding.

Google Trend Analysis

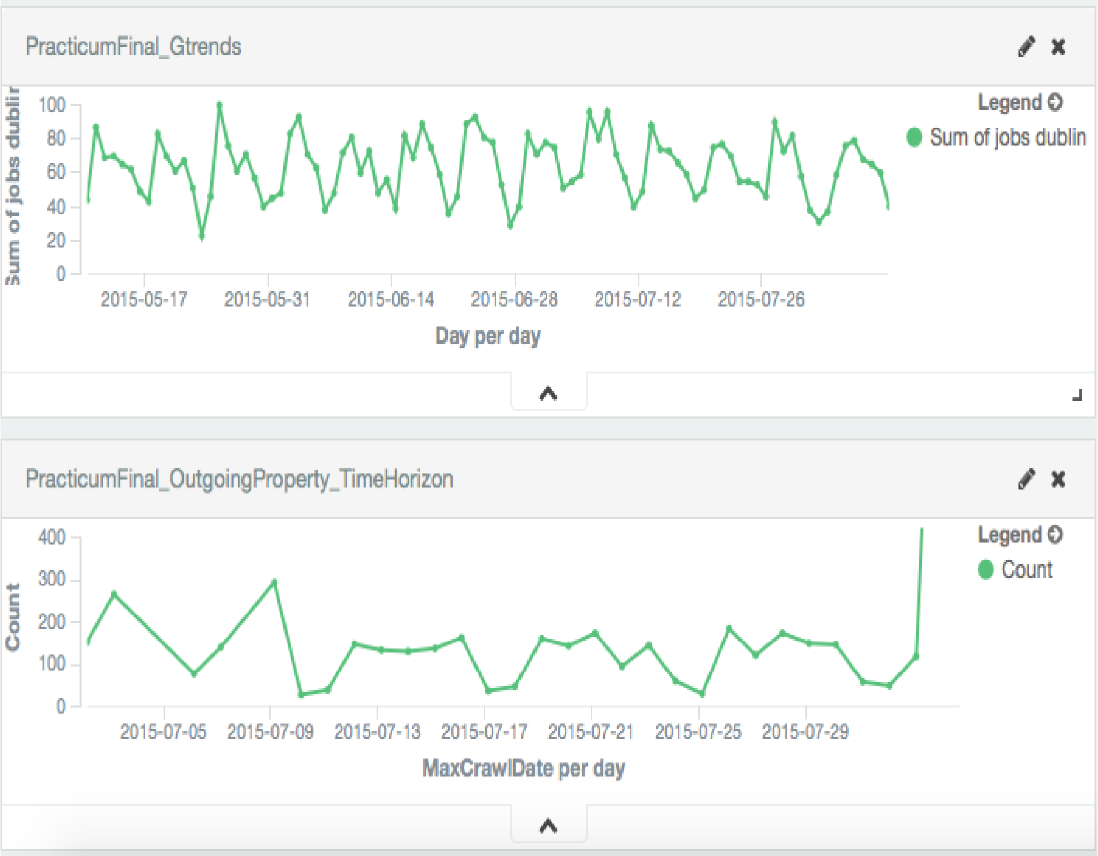


Figure 25: Google trends of 'Jobs in Dublin' and Demand of property

We have taken the count of properties being taken up per week starting from 28 June onwards till 26 July and the job search results from Google Trends for the previous week, previous two weeks and previous three weeks. As we know, there is definitely a lag between getting a job and looking for a property to rent. Once the job is finalized, then one will start finding for a property to rent, which could take weeks. Hence, this was the reason behind collecting the data is that fashion.

WEEK	COUNT OF OUTGOING PROPERTIES	JOB SEARCH		
		Last 1 week	Last 2 weeks	last 3 weeks
28th Jun - 5th July	582	469	918	1356
6th July - 12 July	800	472	935	1391
13th July - 19th July	812	489	961	1424
20th July - 26 July	837	455	944	1416

Table 2: Data from Google Trends search and Number of Outgoing Properties

From the stats, we can infer that it takes almost three weeks for a person to find a new dwelling place as per his job search. Therefore, we won't be wrong in saying that the job postings or the increase in job openings does increase housing demand and ultimately leads to price rise.

PracticumFinal_OutgoingProperty_TimeHorizon

It shows the line chart of the total number of outgoing properties or the number of properties being taken up on weekly basis. It can be altered to see the line chart on daily or monthly basis. This plot can be useful if forecasted for a longer duration of time (quarters or years) which could tell us an appropriate time of putting up a property on market with seller's perspective to make maximum profit. From the current data, we tried to correlate number of outgoing properties with the job market of Dublin as discussed in the previous section.

5.3 Predictive Model

The decision tree has been made using J48 classification algorithm. Random Forest has also been implemented in this work. For better interpretation and the limitation of size, we have used only three postal codes for making this tree i.e. Dublin 4 (highest price), Dublin 10 (longest duration) and Dublin 17 (shortest duration). The Decision Tree is shown below:

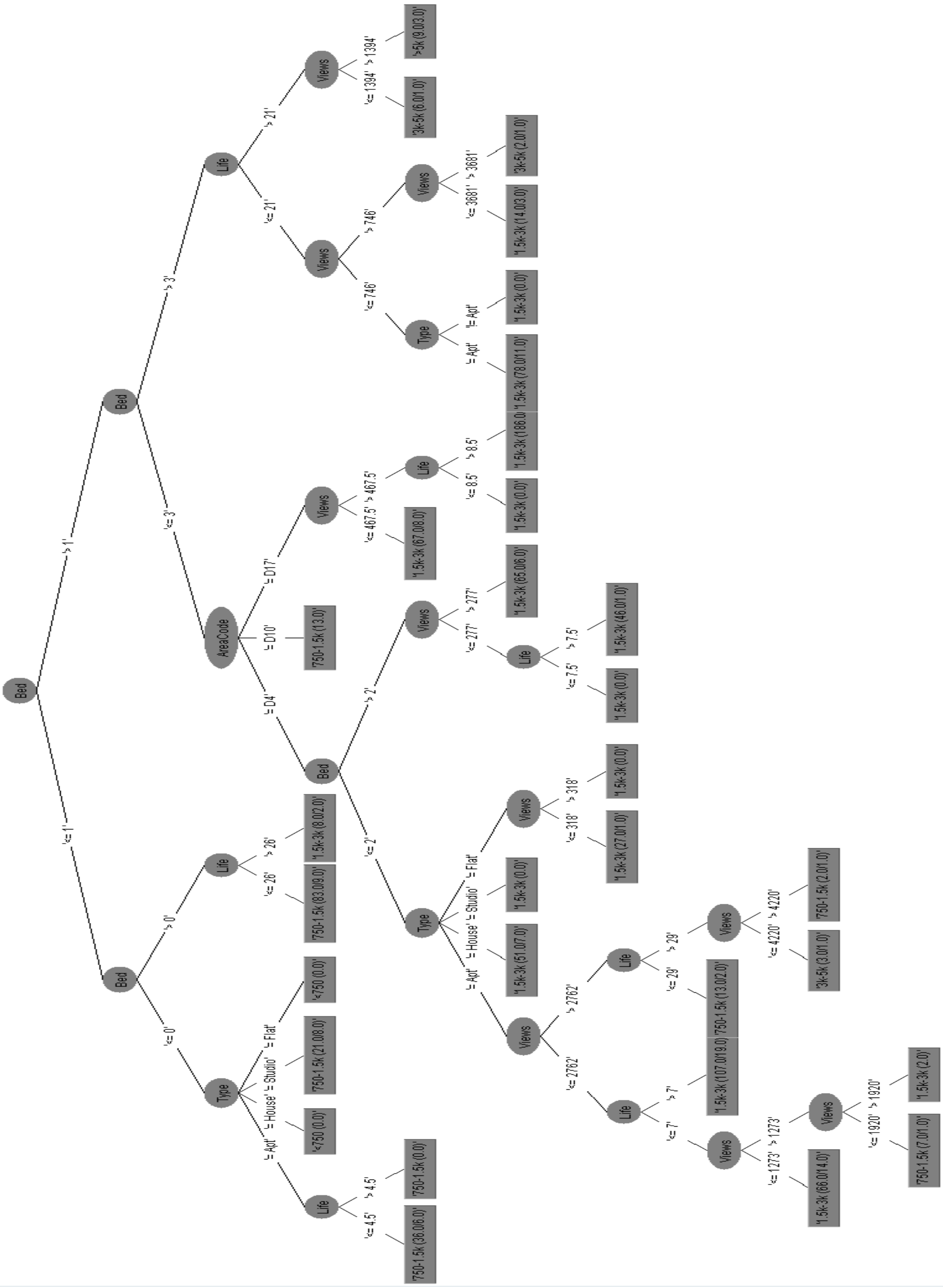


Figure 26: Decision Tree with postal codes Dublin 4, 10 and 17

The class attribute has groups of five rental prices. The root node denotes the number of beds and the terminal node shows the expected rental price of that property. This classification of the rental prices has been done on the basis of 7 other attributes: Life of a Property, Area code, Advertiser Type, Type of a property, Number of beds, Number of Baths and the Number of Views.

This decision tree is useful for all the concerned stakeholders as follows:

- A person looking for a property can have the pre requirements of number of beds, property type and check the appropriate area under his budget or vice versa.
- A property owner can find an appropriate price for his property to let and can also get a brief idea as about in how many days his/her property can be taken up from the market. An investor can have a limited sum of amount to invest on a particular property type then this tree can be helpful to them in determining the approximate life of a property in days to be taken up for the return on investment purpose.
- The real estate can determine which set of attributes of a property will enable them to fetch the price they require.

The statistical measures are shown in Table 3 below for Decision Tree with three postal codes and with all the postal codes. The statistics for Random Forest is also mentioned. The accuracy, precision and recall has been calculated for our multiclass model. Precision denotes the fraction of predicted instances that are relevant and recall is the fraction of relevant instances that are predicted.

Summary	J48 algorithm on 3 postal codes	J48 algorithm on complete data set	Random Forest on complete data set
Correctly Classified Instances (Accuracy)	82.42%	83.77%	71.51%
Precision (weighted Avg.)	82.3%	83.8%	71.31%
Recall (weighted Avg.)	82.4%	83.8%	71.62%

Table 3: Prediction Model Statistics

The Importance Measure Gini Index plot from Random Forest is available in Appendix 8.

Chapter 6: Analysis and Discussion

In this section, detailed analysis has been done in order to meet the needs of the stakeholders by using the trends and patterns from the above observations under the result section.

6.1 From Real Estate Perspective: Forecast the demand for a type of property in an area

With the number of attributes we have currently available for the practicum work, the demand of particular property is determined by its number of views and the rate at which it is being taken up that is the duration of the property. Hence, by combining and correlating various observations the possible demand forecast has been done as below:

<u>Apartment</u>:

Dublin 5 has the highest movement for apartments which includes 1, 2, 3 and 5 bed apartments with average price of €1261 which is not very high as compared to other areas which we believe is one attractive feature for its shorter life. However, the number of properties available are among the least. Hence, more apartment would be needed in this area in near future.

Views in Dublin 24 is very high, however the movement is among the slowest even though they are moderately priced. Possible reason for these behavior could be less number of options to choose from. Apartment can be targeted here as well.

<u>House</u>:

Houses show a good response in Dublin 17. The average life is of 6 days and the lowest compared to other areas. Views and price are also considerably good but the number of available properties is very low which could lead to price rise in that region for Houses as the demand increases. Therefore, Dublin 17 looks good for construction of houses especially three beds.

Flat:

In Dublin 22, flats have seen the highest movement. So Dublin 22 can be considered for construction of more flats. Views are the highest for flat in Dublin 7, 24 respectively. However, the movement in these areas are slow even though being moderately priced possibly because less number of flats available hence people have less option to choose from. Hence, if flats are targeted there are chances of good response in these areas.

Studio:

Average life of studios is shortest in Dublin 13 on an average of 2 days. It has good number of views also the average price is 2nd highest compared to other parts of Dublin. Hence building studios could be profitable. Views for studios are more in Dublin 1, 2, also moderately priced. Dublin 1 and 2 can also be thought upon to build flats as current availability is quite low.

In addition, the prediction model built can help the real estate estimate the price of the property they are planning to build in the area. Also, the Kibana dashboard will allow them to understand the demand for the areas based on the views and life of the properties. The Google trend showing the jobs searches can act as an indication of the upcoming demand for the property in the housing market.

6.2 Property Owners/Investors Perspective: To ease the rent/sell ratio

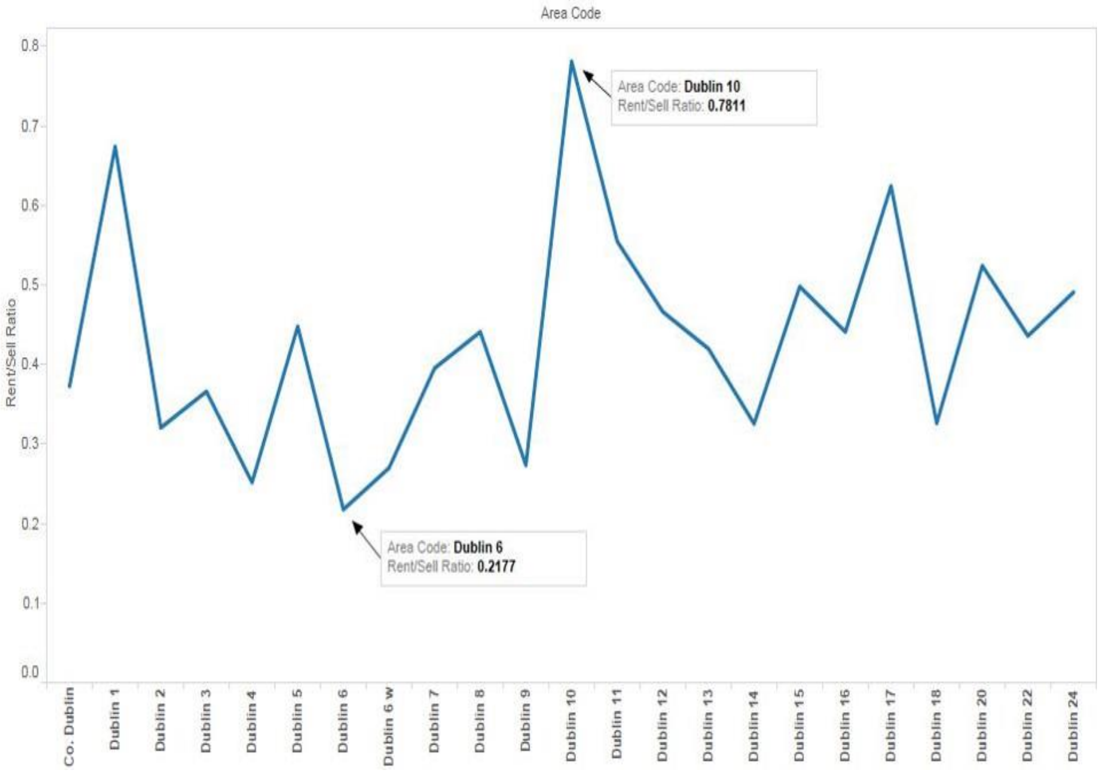


Figure 27: Return on Investment (Rent/Sell ratio) Area Wise

We have plotted a line chart as shown in Figure 27 indicating rent/sell ratio of properties in area wise. It has been observed that the highest rent to sell ratio are in Dublin 10 (Ratio 0.7811) followed by Dublin 1 and 17. It means that if an investor buys a property in these area then he is more likely to get best returns on his investment. However, the areas like Dublin 6, 9, 14 and 18 has the least rent/sell ratio thereby it could be not so good a decision for investing in terms of good returns in these areas.

In addition, the Kibana dashboard gives the views and the life of the property, which is the indication of the demand of that area. Therefore, the property owners would have a fair idea of the fate of their property or can plan ahead before investing in that property.

6.3 Renters Perspective: Providing a Holistic View

Renters willing to rent a house can have an overall view of the prices and other trends across different areas in real-time and then select a place to rent in accordance with other external preferences as they like which is outside the scope of our practicum. For majority of the people the first and foremost factor for renting a house would be subject to its price. Then comes other factors such as the distance from the city or the transport facilities etc. With the help of the dashboard, interested renters can decide on the area to rent by looking at number of beds of their choice and their corresponding prices. Also they can decide on the basis of the property type (apartment/flat/house/studio) offered by the agents as well as the private advertiser and compare their prices to finalize on the area which best suits their interest or can simply select the one with the lowest price.

Therefore, by deciding on the type of property and the number of beds the renter can have a view of the prices for the mentioned combination in one glance and also can see the prices of the properties for the advertiser type and hence can compare and decide the better deal among them. This makes the task of house hunt a lot easier. Also one can see the number of such properties available and the life of the properties in those areas, which could further help in the decision making process of taking the move at the correct time before the property is taken up by somebody else. One can also chose the see the number of views for those properties and know if it’s likely to be taken up by someone else and act then accordingly. The predictive model would be an add-on benefit that could let them know the price of the property they like on the basis of their choice of the attributes.

6.4 Detailed Analysis of the Selected Area Codes

6.4.1 Dublin 10

Dublin 10 has been selected to analyze further as among all the other areas Dublin 10 has the slowest moving property. The life of the properties in Dublin 10 is the highest with average being 15 days.

It has more of 2 bed properties whereas life of 1 bed properties are shorter.1 bed properties are taken up at a faster rate than the 2 bed ones.

One bed houses stay for less duration in the market an average of 2 days but there are less number of one bed houses available. Instead there are more of one bed apartment which apparently has very slow movement of 19 days. The three bed property seems to be the slowest to move out from the market which probably is the main reason for Dublin 10 to be the slowest in terms of the property movement.

Since, one bed houses are popular, studio could also be a good option in Dublin 10 which at present has none.

External factors affecting its movement

Dublin 10 is a center of national commercial distribution business. It has a hub of all the major automobile companies such as Toyota, Nissan, and General Motors etc. Therefore, this could be one reason of its very slow movement of the residential properties as it is generally considered as a commercial place.

6.4.2 Dublin 4

Dublin 4 has been chosen for analysis because it is the area with highest average price of €1976 monthly. The life is probably longer in this area would be because of its very high price. House is contributing for its major price hike. Houses are the most expensive in Dublin 4 from all other areas. Also could be studios as the views is really high and life also less within Dublin 4 which shows it’s in demand but less in count which could be the reason behind high prices.

External Factors responsible for its high price:

One reason could be its geographical location and is well connectivity with all the means of transport. It is where all the embassies of Dublin are located so many diplomatic residences are located in Dublin 4. In addition, it’s considered one of the safest place to be in. The other few center of attractions in this area are Royal Dublin Society (RDS), Lansdowne Road Ground and premium hotels.

Chapter 7: Measure of Success

7.1 Survey

The complete model was distributed with full access to the following three profiles:

- Real Estate Agent
- Working Professional
- International Student

Following are the set of questions:

1. Is it interpretable?
2. Does it ease the initial decision making process?
3. Does it give a quick overview of the housing market?
4. Does it help you to find an area of your choice?
5. Does it give you an approximate idea of the demand?

The complete questionnaire form is available in the appendix section 6.

7.1.1 Feedback

Real-Estate Agent

The agent agreed to the results that were produced by our model. The trends and patterns that were observed was in sync with his knowledge of the market. The dashboard seemed extremely handy to throw a glance and have a quick update. From a real estate perspective, he now knows which area is viewed more and what the movement of the property is area wise, which he was previously unaware. He said this could be one of the potential measures of the demand and hence of good importance. In addition, the number of the properties area wise on the heat-map provides a very good insight, which would help them plan their forthcoming projects.

International Student

An incoming UCD student was surveyed. He had no clue about which area to choose from. He only had number of bed in his mind. This dashboard helped him narrow down few areas such as Dublin 6w and Dublin 7 based on his preferred choice of 1 bed property and proximity to UCD which best suited his budget (<1000€). Then with help of other external sources, he concentrated on those areas in particular to find a property.

Working Professional

A working professional from Accenture was looking for a property to rent nearby to her office. She was not sure whether to look for an apartment or a house. Our dashboard helped her to compare the property prices for both of the property types. Additionally, suggested the prices put up by the advertiser type. She went ahead with the house in Dublin 1 put up by private advertiser instead of going to the agents.

7.2 Statistical Analysis

As part of the statistical validation, the output from the Random Forest is used. The model is 71.51% accurate and for class A, 77.83% of the predicted instances are relevant. In case of the recall measure for class A, 83% of the relevant instances are retrieved and likewise for other classes. In order to combine the two stats, F_1 score has been calculated which evenly weighs recall and precision as shown in Figure 28.

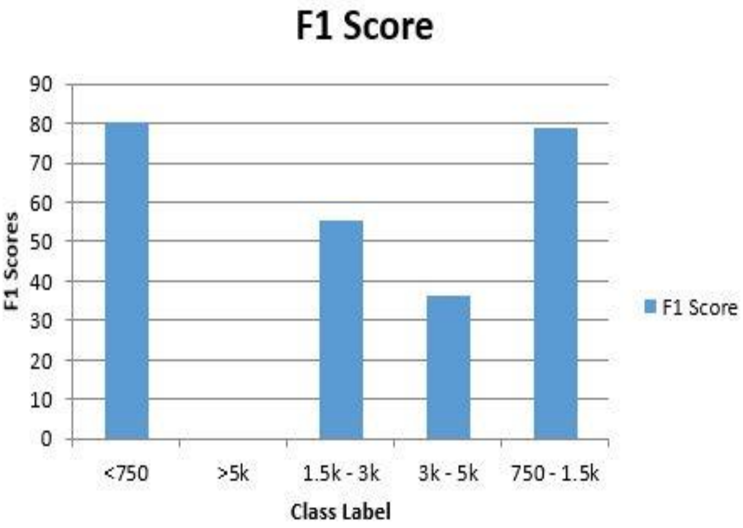


Figure 28: F1 Score from Random Forest

However, class B (>5k) is misclassified as the data itself is very less in that range. Rest other classes shows good F_1 scores.

Chapter 8: Learnings

During the course of our work, we have realized the power of web crawlers. As we know Internet has wide expanse of information, to get the relevant information out of it is a challenging task. This is where web crawlers play an effective role. Likewise, in our work as well, we crawled the information from Daft's web site, which has thousands of listings on it. Crawlers helped us to retrieve the relevant information from all the listings and create our dataset.

In addition, we had a good hands-on experience on industry standard analytic tools such as Tableau, Elasticsearch and Kibana, which enabled us to build our near real-time dashboard. This is important because in this dynamic environment where information in the form of data is changing so rapidly, it is the utmost need of the hour to be at pace with the changing environment and have knowledge of the happenings. The real-time dashboard helps the users to be updated with all the changes, which helps them to be well informed and participate in the business accordingly. Tableau helped in analyzing the hidden patterns and trends which otherwise would have remained unknown.

While creating the prediction model, we came across various R packages such as rpart, random forest, etc. which is helpful when the observations are less but attributes are many, which was true in our case.

Chapter 9: Conclusion

The aim of the practicum has been successfully achieved. The questions of the main stakeholders categorized above have been answered effectively. The report in the form of dashboard that has been developed using Kibana would provide near real-time update on the trends of the housing market like the price, the amount of properties available, the duration of the property in certain areas, etc. which would cater to the demands and the needs of the stakeholders. The report is definitely of great help as the current reports that are being generated from a number of sources are usually quarterly, which serves no much purpose to the renter /buyer as it is just an analysis of the previous quarters. However, the report that is generated as part of this practicum is focused to assist the stakeholders in their respective decision-making in their area of concern. It is advantageous over the usual reports as this is real-time so one can get an immediate idea or develop a foundation of the market before taking the next step. If the data set is increased and taken over months and years, the patterns and the trends will be more clearly visible and provide more insights. Furthermore, the accuracy of the prediction model would increase. Nevertheless, from the viewpoint of one month's data these results are of good significance.

As advised by IBM Research, the whole implementation of this practicum is shared in GitHub account so that it can act as a reference or a foundation for anyone who would like to work on this topic further. It includes the data set, codes, scripts, methods and a detailed instruction in a form of 'read me' file that contains the complete in depth procedure starting from crawling of data to indexing it into Elasticsearch for visualization in Kibana's dashboard.

Chapter 10: Future Work

Due to time constraint, other external factors could not be considered for dissection of the market. Since, it was one month of the data that was collected, many of the factors could not be possibly correlated like those of the incoming international students, economic growth because these stats are either quarterly or yearly. Some are more of seasonal trend, therefore these could not be mapped to the monthly data.

Other factors that can be considered are the distance from the city, transportation, quality of the uploaded images, price history, etc. The price history details were collected but given the duration of the practicum, significant insights could not be drawn out of it. To limit the number of different tools used, under certain conditions, the data parser script written in Python can be modified to do much of the other tasks. For example, the retrieval of unique IDs that was done in SQL can be implemented in Python to make the whole practicum drifted towards more automation side.

At the moment, the Kibana dashboard is hosted on local server but it can also be hosted on a web server and the read only access can be given to the general public. However, the streaming of data into Elasticsearch can be done at the backend by necessary administrator rights.

Another potential work that can be done is to build a data driven recommender system that could be customized based on the requirements. The practicum can be extended by including other area in Ireland and not just the capital city.

Appendix

Appendix 1: Data Formatting Script

The data formatting script ‘**DaftDataParser.py’** is available in an enclosed DVD and it uses the following modules:

- JSON: This module is used to process the JSON files
- Dateutil.parser: This module is used to parse and process dates
- Geocoder: for getting geo-coordinates
- Os: This module is used for interacting with the operating system
- Sys: This module provides access to some variables used or maintained by the interpreter and to functions that interact strongly with the interpreter.

Directory Setup: The script first checks the existence of user specified folder directory path. If it does not find, then it will create one with the name “OutputJSON” to store all the processed JSONs.

Processing JSON: It then searches for all the files in the user specified directory which has *.JSON file type. In case it finds these files, then it will treat it as input file and set its permission as ‘read only’ and simultaneously it makes the output file with same name inside ‘OutputJSON’ folder and sets its permission as ‘write’ for the further processing.

Initialisation of Dictionary: The script then save all the values for each key from the input data JSON to the new variable called ‘all_listings’. It also filtered out the empty results at this step. The rest of the code works in a loop for all the property listings in ‘all_listings’ one by one. Inside this loop, the dictionary has been declared to save all the keys for each listings. This dictionary was used as an “associative arrays” indexed by keys. It first fetches out the crawled date of a particular listing and saves it in a the dictionary as ‘filtered_data[‘crawlDate’]’.

Property Type: It extracts the key ‘type’ which contains the number of beds, number of baths and property type (Apartment/Flat/House/Studio Apartment) and all these fields are separated by ‘ | ’. In order to extract all these fields separated, the script conducts the search on the key ‘type’ and it splits its value with ‘ | ‘. In Daft, the number of beds are listed first followed by number of baths. Hence, the script saves the first value before the split operator as number of beds and the second as number of baths. If it detects the type as “Studio Apartment” then it automatically sets the number of beds and baths as ZERO.

PageURL and urlID: The script then saves the listing url by taking value from key ‘_pageUrl’. For better search results and to handle search queries for efficiently, the script assigns a unique urlID to each property listing. This unique urlID is fetched from the end of original URL as it contains a unique code associated with each listing. The uniqueness has been verified before implementing it into the script. The script searches for an operator ‘/’ from the end and saves the complete number just after it as unique ID into ‘filtered_data[urlID]’.

Property Price: The original JSON has price listed in € along with payment duration (Monthly/Weekly) in STRING type. The script first removes the currency symbol € and any other junk characters from the price and splits it into two fields: ‘filtered_data['price']’ takes the price and converts it into INTEGER type and ‘filtered_data[‘paymentDuration’] takes has the payment duration and saves it as STRING type.

Extraction of Dates: The original data set consists of three dates: 1) ‘date/_utc’ is the advertised date when the advertiser listed a property. 2) ‘date/_source’ is a system generated date which is saved as sourceDate in the dictionary. 3) ‘date’ is the actual crawled date which is saved as ‘unixAdDate’ in the dictionary.

Geo-coordinates and Postal Codes: The script uses open source Google’s Geocoding Services to determine the geographical coordinates of each property listing. Geocoding Services is triggered by calling the function ‘geocoder.Google()’ on each listing’s ‘area’ key in the original JSON. As discussed earlier, ‘area’ key has the complete address of a property. The return type of this function is ‘latlng’ which means that it gives the latitude of a property first followed by its longitude. However, the Kibana has the limitation in mapping the coordinates as it only accepts the longitude value first followed by the latitude. In order to do it, the script swaps the coordinates and saves the swapped values of geographical coordinates in ‘filtered_data['coordinates']’.

The postal codes were fetched from the key ‘area’ using a search split on ‘, ‘starting from the last as it was verified that the postal codes were captured at the end of each property’s address. The postal codes were saved in ‘filtered_data['postalCode']’.

Advertiser Details: The script saves the advertiser details from the source dictionary and adds it to the filtered dictionary under ‘filtered_data['advertiser']’.

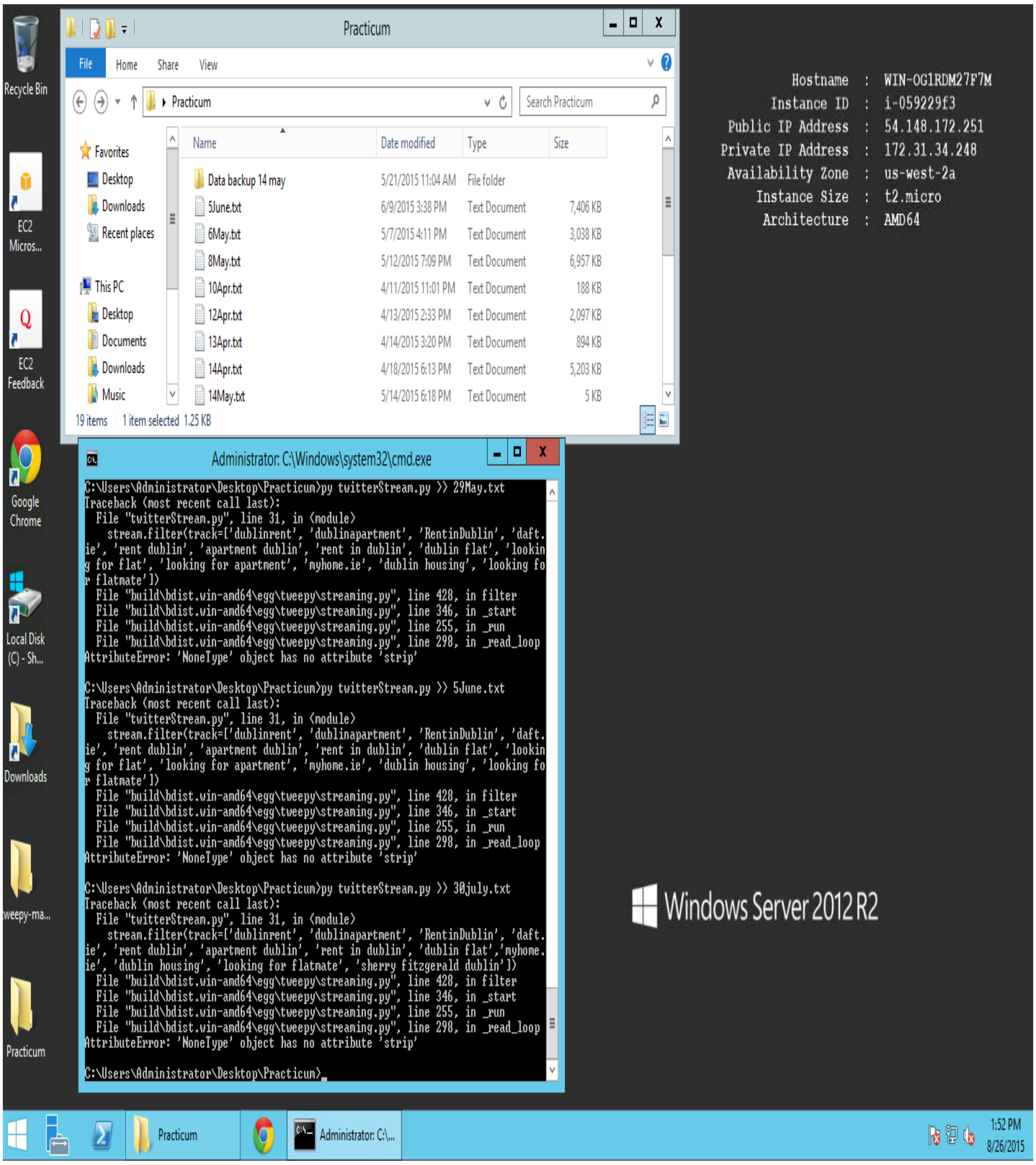
Pricing History: The script saves the all the pricing history of a particular listing by applying a split search on the currency symbol €. It also saves the first listing by searching a keyword ‘First Listed’ and saves all the further values of price changes in the ‘changed_price’ and the ‘updated_amount’ keys along with their respective dates in pricing_change_log['date'].

Photos and Views: The script extracts the number of photos and saves it into ‘filtered_data['photos']’. The source data from the crawler captured the number of views in a STRING type. Therefore, the script is written in such a way that it first extracts the number of views and then it converts it into integer type and saves it into filtered_data['views'].

Finally, all the above filtered results are added to the result list as ‘filtered_result_JSON’ and at this point, the script closes the output file. The parsed JSON files for each crawled dates are then created inside the ‘OutputJSON’ folder. The script is available in an enclosed DVD.

Appendix 2: Amazon Web Services

We have setup an account on Amazon Web Services (AWS) to run our twitter script. We have enrolled a student 90 days' free subscription from Amazon to setup this account. Following is a screenshot of this account showing a twitter script running along with the folder in which data was saved:



Name	Date modified	Type	Size
Data backup 14 may	5/21/2015 11:04 AM	File folder	
5June.txt	6/9/2015 3:38 PM	Text Document	7,406 KB
6May.txt	5/7/2015 4:11 PM	Text Document	3,038 KB
8May.txt	5/12/2015 7:09 PM	Text Document	6,957 KB
10Apr.txt	4/11/2015 11:01 PM	Text Document	188 KB
12Apr.txt	4/13/2015 2:33 PM	Text Document	2,097 KB
13Apr.txt	4/14/2015 3:20 PM	Text Document	894 KB
14Apr.txt	4/18/2015 6:13 PM	Text Document	5,203 KB
14May.txt	5/14/2015 6:18 PM	Text Document	5 KB

```
C:\Users\Administrator\Desktop\Practicum>py twitterStream.py >> 29May.txt
Traceback (most recent call last):
  File "twitterStream.py", line 31, in <module>
    stream.filter(track=['dublinrent', 'dublinapartment', 'RentinDublin', 'daft.
ie', 'rent dublin', 'apartment dublin', 'rent in dublin', 'dublin flat', 'lookin
g for flat', 'looking for apartment', 'myhome.ie', 'dublin housing', 'looking fo
r flatmate'])
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 428, in filter
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 346, in _start
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 255, in _run
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 298, in _read_loop
AttributeError: 'NoneType' object has no attribute 'strip'

C:\Users\Administrator\Desktop\Practicum>py twitterStream.py >> 5June.txt
Traceback (most recent call last):
  File "twitterStream.py", line 31, in <module>
    stream.filter(track=['dublinrent', 'dublinapartment', 'RentinDublin', 'daft.
ie', 'rent dublin', 'apartment dublin', 'rent in dublin', 'dublin flat', 'lookin
g for flat', 'looking for apartment', 'myhome.ie', 'dublin housing', 'looking fo
r flatmate'])
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 428, in filter
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 346, in _start
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 255, in _run
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 298, in _read_loop
AttributeError: 'NoneType' object has no attribute 'strip'

C:\Users\Administrator\Desktop\Practicum>py twitterStream.py >> 30july.txt
Traceback (most recent call last):
  File "twitterStream.py", line 31, in <module>
    stream.filter(track=['dublinrent', 'dublinapartment', 'RentinDublin', 'daft.
ie', 'rent dublin', 'apartment dublin', 'rent in dublin', 'dublin flat','myhome.
ie', 'dublin housing', 'looking for flatmate', 'sherry fitzgerald dublin'])
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 428, in filter
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 346, in _start
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 255, in _run
  File "build\bdist.win-amd64\egg\tweepy\streaming.py", line 298, in _read_loop
AttributeError: 'NoneType' object has no attribute 'strip'

C:\Users\Administrator\Desktop\Practicum>_
```

```
         Hostname  :  WIN-OG1RDM27F7M
      Instance ID  :  i-059229f3
Public IP Address  :  54.148.172.251
Private IP Address :  172.31.34.248
Availability Zone  :  us-west-2a
    Instance Size  :  t2.micro
     Architecture  :  AMD64
```

Appendix 3: Twitter Script

The python script for twitter '**TwitterStream.py**' helped us in retrieving the tweets for the keywords entered. First it imports the necessary methods from the tweepy library. Then with the help of authentication details entered like the CONSUMER_KEY, CONSUMER_SECRET, OATH_TOKEN etc. it connects to the twitter streaming API once the authentication is validated. Then it filters the twitter streams to capture the data by the keywords that has been entered. The script is available in an enclosed DVD.

Following is an example of one of the tweets captured by this script for 30th July 2015:

```
{"created_at":"Thu Jul 30 14:30:12 +0000
2015","id":626761717683519489,"id_str":"626761717683519489","text":"Caterina is looking for a
#Flat in #Dublin for 181 nights from 01 Sep 2015 for  \u20ac600 p\/m. Make an offer now:
http:\/\/t.co\/lx5Boonhwg #O","source":"\u003ca href=\"http:\/\/flatclub.com\"
rel=\"nofollow\"\u003eFlatClub\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_re
ply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_scr
een_name":null,"user":{"id":2396321714,"id_str":"2396321714","name":"Live
Demand","screen_name":"FCLiveDemand","location":"London","url":"https:\/\/flat-
club.com\/livedemand","description":"LiveDemand is a reverse marketplace, allowing hosts to browse
freshly submitted guest requests and send them an offer. Best of all, it is
free.","protected":false,"verified":false,"followers_count":13,"friends_count":4,"listed_count":10,"fav
ourites_count":0,"statuses_count":4378,"created_at":"Tue Mar 18 14:40:09 +0000
2014","utc_offset":3600,"time_zone":"London","geo_enabled":false,"lang":"en-
gb","contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED","profile
_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_backgro
und_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_backgroun
d_tile":false,"profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED","profile_side
bar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"prof
ile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/531785447265951745\/ynhd-
Kph_normal.jpeg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/53178544726
5951745\/ynhd-
Kph_normal.jpeg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/2396321714\/141
5623982","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent"
:null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null}
```

Appendix 4: LinkedIn Script

The python script for LinkedIn ‘**LinkedInScript.py**’ first authenticates the user credentials by making a secured connection with LinkedIn developer server with the entered CONSUMER_KEY, CONSUMER_SECRET, USER_TOKEN and USER_SECRET. It then takes the company names from the user and returns the number of job opening by them with other details. But it couldn’t determine the number of applied applications and their geographical locations. The script is available in an enclosed DVD.

Appendix 5: SQL Script

The following SQL script is used to determine the life of a property:

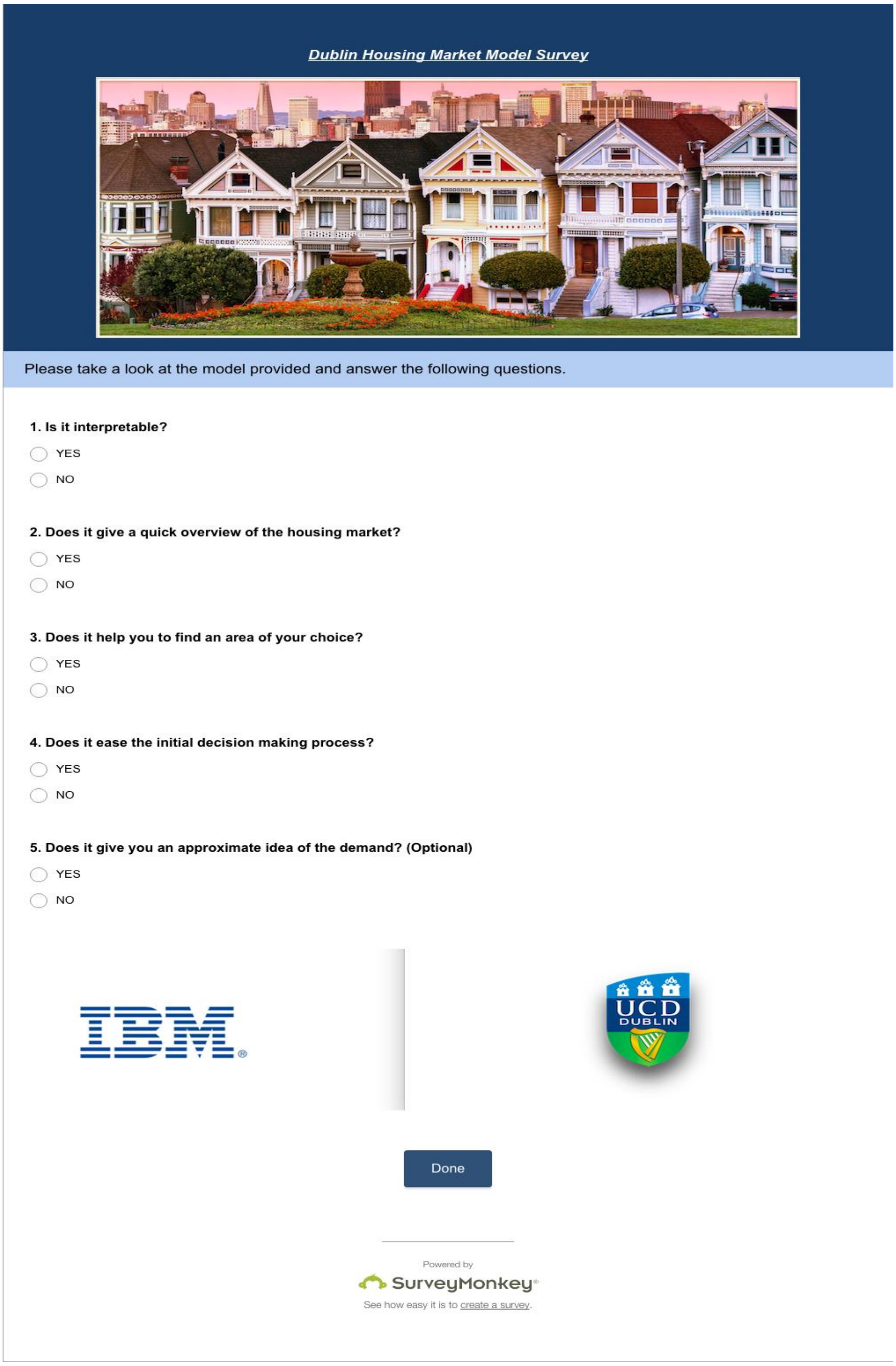
```
SELECT [property_transaction_master].[Urlid], [property_transaction_master].[Area
Code],MIN([property_transaction_master].[AdvertisedDate]),MAX([property_transaction_master].[
Crawl Date]) AS [Max Crawl Date],MAX([property_transaction_master].[Crawl Date])-
MIN([property_transaction_master].[Advertised Date]) as diff

FROM property_transaction_master

GROUP BY [property_transaction_master].[Urlid],[property_transaction_master].[Area Code];
```

The whole dataset is stored in the main table property_transaction_master created in MS Access. Each property appears multiple times in the dataset until it is removed from the site which means it has been taken away from the market. This SQL select query retrieves the maximum crawled date and minimum advertised date for each unique property id from the whole dataset. The difference between dates are calculated to find the duration of the property for which it was advertised on the site and gives their corresponding area code.

Appendix 6: Questionnaire Form



Dublin Housing Market Model Survey

Please take a look at the model provided and answer the following questions.

1. Is it interpretable?

○ YES

○ NO

2. Does it give a quick overview of the housing market?

○ YES

○ NO

3. Does it help you to find an area of your choice?

○ YES

○ NO

4. Does it ease the initial decision making process?

○ YES

○ NO

5. Does it give you an approximate idea of the demand? (Optional)

○ YES

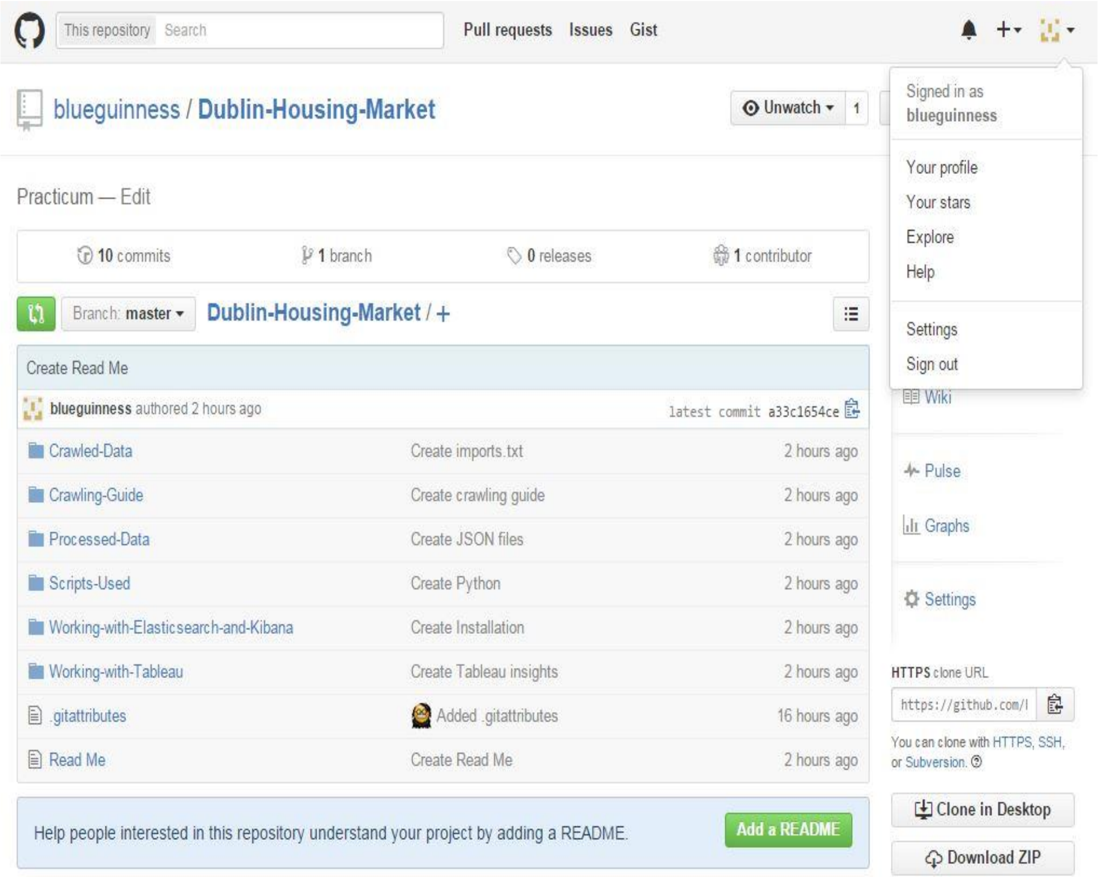
○ NO

Powered by SurveyMonkey

See how easy it is to create a survey.

Appendix 7: GitHub Repository

As advised by IBM Research, we have shared our working, data sets, complete script along with a detailed ‘Read Me’ file on GitHub repository, which can be used as reference for future work on this topic.



blueguinness / Dublin-Housing-Market

Practicum — Edit

10 commits · 1 branch · 0 releases · 1 contributor

Branch: master · Dublin-Housing-Market / +

Create Read Me

blueguinness authored 2 hours ago · latest commit a33c1654ce

Crawled-Data	Create imports.txt	2 hours ago
Crawling-Guide	Create crawling guide	2 hours ago
Processed-Data	Create JSON files	2 hours ago
Scripts-Used	Create Python	2 hours ago
Working-with-Elasticsearch-and-Kibana	Create Installation	2 hours ago
Working-with-Tableau	Create Tableau insights	2 hours ago
.gitattributes	Added .gitattributes	16 hours ago
Read Me	Create Read Me	2 hours ago

Help people interested in this repository understand your project by adding a README.

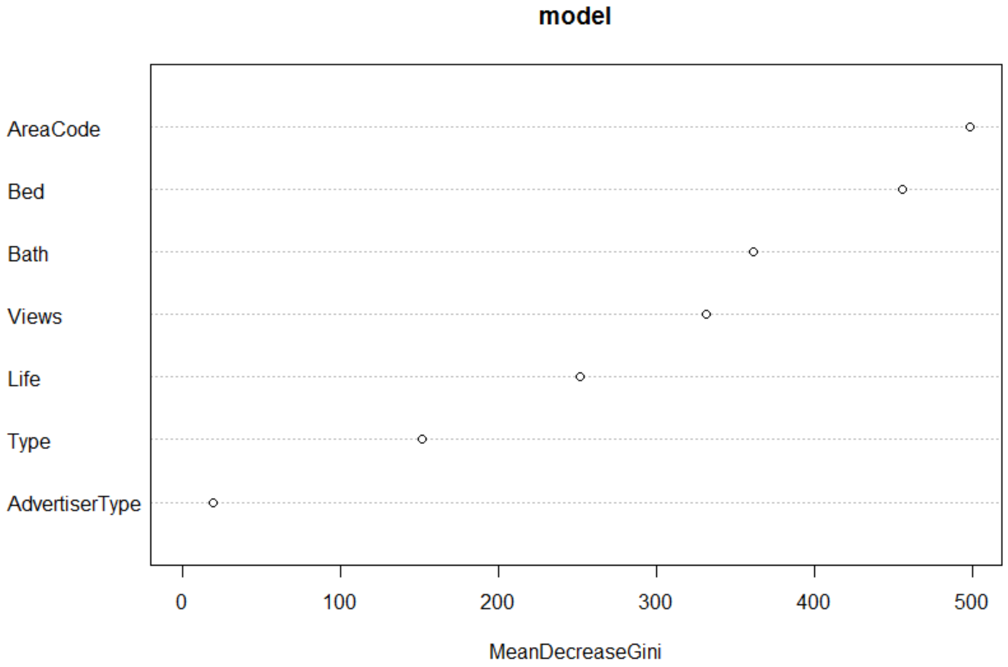
HTTPS clone URL

https://github.com/l

You can clone with HTTPS, SSH, or Subversion.

Appendix 8: Random Forest Importance Measures Model

Following is an importance measure plot (Gini Index) generated by Randon Forest package that we have used on the complete data set.



References

Cho, J. (2001). *CRAWLING THE WEB: DISCOVERY AND MAINTENANCE OF LARGE-SCALE WEB DATA*. 1st ed. [ebook] stanford. Available at: http://oak.cs.ucla.edu/~cho/papers/cho-thesis.pdf [Accessed 24 Jun. 2015].

Choi, H. and Varian, H. (2011). *Predicting the Present with Google Trends*. 1st ed. [ebook] Available at: http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf [Accessed 23 Jul. 2015].

Chukwugozie Nsofor, G. (2006). *A Comparative Analysis Of Predictive Data-Mining Techniques*. 1st ed. [ebook] Tennessee, Knoxville. Available at: http://web.utk.edu/~xli27/rawDocs/XLI/Godswill's%20Thesis%2006-19-06ok.pdf [Accessed 10 Jun. 2015].

Corelogic.com.au, (2015). *Data & Analytics - Residential Property Insights - Data & Analytics - Residential Property Insights| CoreLogic RP Data*. [online] Available at: http://www.corelogic.com.au/service/data-analytics [Accessed 6 Aug. 2015].

D. Correa, C., Chan, Y. and Ma, K. (2008). *A Framework for Uncertainty-Aware Visual Analytics*. 1st ed. [ebook] California: University of California at Davis. Available at: http://vis.cs.ucdavis.edu/papers/vast09.pdf [Accessed 12 Jul. 2015].

Daft.ie (2015). *Chronic supply shortages persist in the rental market*. [Online] Available at: https://www.Daft.ie/report/ronan-lyons-2015q1-rental [Accessed 1 July 2015].

Elastic.co (2015). *Elasticsearch | Search and Analyze Data in Real Time*. [Online] Available at: https://www.elastic.co/products/elasticsearch [Accessed 10 June 2015].

Elastic.co (2015). *Kibana | Explore and Visualize Your Data*. [Online] Available at: https://www.elastic.co/products/kibana [Accessed 10 July 2015].

Fayyad, U .M., Piatetsky-Sharpio, G. Smyth. P. and Uthurusany, R. (1996). *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence Menlo Park, CA, USA. [Accessed 17 July 2015].

FEW, S. (2007). *DATA VISUALIZATION PAST, PRESENT, AND FUTURE*. 1st ed. [ebook] Cognos Innovation Centre. Available at: http://www.perceptualedge.com/articles/Whitepapers/Data_Visualization.pdf [Accessed 10 Jul. 2015].

Frawley, W. J, Piatetsky-Shapiro, G. and Matheus, C. J. (1991). *Knowledge Discovery in Databases: An Overview Knowledge Discovery in Databases*, Pg. 1 - 27, AAAI/MIT Press. [Accessed 17 July 2015].

Golfarelli, M. & Rizzi, S. (2009). *Data Warehouse Design : Modern Principles and Methodologies*. McGraw-Hill Osburn. [Accessed 17 July 2015].

Guide, G. (2015). *Irish house prices surging!*. [online] Global Property Guide. Available at: http://www.globalpropertyguide.com/Europe/ireland/Price-History [Accessed 1 Aug. 2015].

Gupta, A. and Dubey, G. (2012). *Identifying Buying Preferences of Customers in RealEstate Industry Using Data Mining Techniques*. 2nd ed. [ebook] Amity University, Noida. Available at: http://www.academia.edu/2304462/Identifying_Buying_Preferences_of_Customers_in_Real_Estate_Industry_Using_Data_Mining_Techniques [Accessed 8 Aug. 2015].

Independent.ie, (2014). *Ten steps to solving Dublin's housing crisis - Independent.ie*. [online] Available at: http://www.independent.ie/life/home-garden/homes/ten-steps-to-solving-dublins-housing-crisis-30111454.html [Accessed 20 Jul. 2015].

Independent.ie, (2015). *Rental market movement: Dublin and Leinster - Independent.ie*. [online] Available at: http://www.independent.ie/irish-news/the-rent-report/rental-market-movement-dublin-and-leinster-31302183.html [Accessed 3 Jul. 2015].

Kimball, R. et al. (2008). *The Data Warehouse Lifecycle Toolkit*. 2nd Ed. Wiley. [Accessed 17 July 2015].

Liaw, A. and Wiener, M. (2002). *Classification and Regression by randomForest*. 2nd ed. [ebook] Available at: http://www.bios.unc.edu/~dzeng/BIOS740/randomforest.pdf [Accessed 17 Aug. 2015].

Lyons, R. (2015). *An analysis of recent trends in the Irish rental market for 2015 Q1*. 1st ed. [ebook] Ireland: Daft.ie. Available at: http://www.daft.ie/report/q1-2015-daft-rental-report.pdf [Accessed 11 Jul. 2015].

Lyons, R. (2015). *The Daft.ie Rental Report*. 1st ed. Ireland. Available at: https://www.Daft.ie/report/q1-2015-Daft-rental-report.pdf [Accessed 14 Jul. 2015].

Mac Coille, C. (2015). *MyHome Property Report, Q2 2015: House price inflation continues to moderate*. 1st ed. [ebook] Dublin: Davy Research. Available at: https://www.davy.ie/research/public/printPdf.htm?id=MyHomeQ22015_30062015.htm [Accessed 1 Jul. 2015].

Maimon, Oded. And Rokach, Lior. (2010). *Data Mining and Knowledge Discovery Handbook*. 2nd ed. [ebook] Springer. Available at: http://www.cs.bme.hu/nagyadat/Data_Mining_and_Knowledge_Discovery.pdf [Accessed 14 Jul. 2015]

Needtagger.com, (2015). *How To Find Real Estate Leads on Twitter*. [online] Available at: http://www.needtagger.com/how-to-find-real-estate-leads-on-twitter/ [Accessed 16 Jun. 2015]

O'Donovan, C. (2015). *Annual Residential Property Review & Outlook*. 1st ed. [ebook] Ireland. Available at: https://www.scsi.ie/media_centre/scsi_annual_residential_property_review_outlook_2015_ [Accessed 29 Jul. 2015]

Olston, C. and Najork, M. (2010). *Web Crawling*. 4th ed. [ebook] Sunnyvale, CA: Foundations and TrendsR in Information Retrieval. Available at: http://infolab.stanford.edu/~olston/publications/crawling_survey.pdf [Accessed 28 Jun. 2015].

Pollak, S. (2014). *Rents soar across Dublin as students begin housing search*. [online] The Irish Times. Available at: http://www.irishtimes.com/news/education/rents-soar-across-dublin-as-students-begin-housing-search-1.1900814 [Accessed 25 Jul. 2015].

Rys, M. and Fai Yau, K. (1997). *Data Extraction from Dynamic Web Sites: Combining Crawling and Extraction*. 1st ed. [ebook] Stanford. Available at: http://infolab.stanford.edu/~rys/papers/crawl.pdf [Accessed 17 Jun. 2015].

Shkapenyuk, V. and Suel, T. (2001). *Design and Implementation of a High-Performance Distributed Web Crawler*. 1st ed. [ebook] Brooklyn, NY. Available at: http://cis.poly.edu/suel/papers/crawl.pdf [Accessed 25 Jun. 2015].

Shoaib, M., Farooqui, M. and Zunnun Khan, M. (2015). *Discovering Web through Crawler: A Review*. 1st ed. [ebook] Aligarh,India: Mangalayatan University. Available at: http://www.academia.edu/9464890/Discovering_Web_through_Crawler_A_Review [Accessed 17 Jul. 2015].

Sun, G., Liang, R., Wu, F. and Qu, H. (2013). *A Web-based visual analytics system for real estate data*. 5th ed. [ebook] China: Science China Information Sciences. Available at: http://link.springer.com/article/10.1007%2Fs11432-013-4830-9#page-1 [Accessed 23 Jul. 2015].

Uğur, A. (2015). *BPFI Housing Market Monitor*. 1st ed. [ebook] Ireland: Banking and Payment Federation. Available at: http://www.bpfi.ie/wp-content/uploads/2015/06/BPFI-HMM-Q1-2015-FINAL.pdf [Accessed 2 Aug. 2015].

Venkat Raman, V., Vijay, S. and Banu K, S. (2014). *Identifying Customer Interest in Real Estate Using Data Mining Techniques*. 1st ed. [ebook] Tamil Nadu,India: International Journal of Computer Science and InformationTechnologies.Available at:http://www.ijcsit.com/docs/Volume%205/vol5issue03/ijcsit2014050389.pdf [Accessed 21 Jul. 2015].