**Une image contenant texte, clipart

Description générée automatiquement**

|  |
| --- |
| **Big Data Mining for Smart Cities: Real Estate Price Prediction using Time-Series Analysis** |

**Student Name**

|  |  |
| --- | --- |
| Main academic Supervisor: | Prof. Christos Tjortjis , International Hellenic University |
| Academic co-supervisor:  Partner supervisor: | Title, First name, LAST NAME, SMACCs Univ.  Title, First name, LAST NAME, affiliation |

**A Master Thesis submitted for the Erasmus Mundus Joint Master Degree on Smart Cities and Communities (SMACCs)**

June 2021

University of Mons, Heriot Watt University, International Hellenic University, University of the Basque Country

|  |  |  |  |
| --- | --- | --- | --- |
| Photos, logos, tournages - Université de Mons | Une image contenant texte, signe, arts de la table, vaisselle  Description générée automatiquement | MA in the Classical Archaeology and the Ancient History of Macedonia,  Thessaloniki, Greece 2021 |  |

Acknowledgements

Abstract

Here goes a summary of the dissertation (1 page max): describes briefly the motivation and objectives, the employed methods and the main findings and conclusions.

This dissertation was written as a part of the MSc in Smart Cities and Communities (SMACCs) under the European Unions Erasmus + Program

**Keywords:** Price Forecast, Multivariate Time Series, Vector Autoregressive Moving Average Models (VARMA), Stationary Time Series,

Syuqran Naim Shamsuddin

June 2021

Table of Contents

[Acknowledgements iii](#_Toc72522490)

[Abstract v](#_Toc72522491)

[Table of Contents vii](#_Toc72522492)

[1 Introduction 1](#_Toc72522493)

[2 Smart Cities 3](#_Toc72522494)

[2.1 Smart City Definition 3](#_Toc72522495)

[2.2 Smart City Infrastructure and Applications 6](#_Toc72522496)

[2.3 Smart Cities in Context 8](#_Toc72522497)

[3 Big Data, Data Mining in Smart Cities 10](#_Toc72522498)

[3.1 Big Data in Smart Cities 10](#_Toc72522499)

[3.1.1 General Concepts 10](#_Toc72522500)

[3.1.2 Big Data Value Chain Operators: 12](#_Toc72522501)

[3.1.3 Platforms 13](#_Toc72522502)

[3.2 Data Mining 14](#_Toc72522503)

[3.2.1 Data Mining functionalities. 14](#_Toc72522504)

[3.2.2 Evaluation and Metrics in Data Mining 16](#_Toc72522505)

[3.2.3 Challenges in Data Mining 16](#_Toc72522506)

[4 Real Estate & Smart City 17](#_Toc72522507)

[4.1 Impact of Smart City in Real Estate 17](#_Toc72522508)

[4.2 Data mining in Real Estate 19](#_Toc72522509)

[4.3 Selected Real Estate Market 20](#_Toc72522510)

[5 Methodology 21](#_Toc72522511)

[5.1 Data Extraction 22](#_Toc72522512)

[5.2 Data Pre processing and Transformation 24](#_Toc72522513)

[5.3 Initial Model Selection 24](#_Toc72522514)

[6 Conclusion and future work 25](#_Toc72522515)

[Bibliography 27](#_Toc72522516)

[Appendix 31](#_Toc72522517)

**List of Figures**

[Figure 1 Smart City Ecosystem Framework [5] 9](#_Toc69911017)

[Figure 2 Implementation Strategies in Smart Cities 15](https://heriotwatt-my.sharepoint.com/personal/sb228_hw_ac_uk/Documents/WF_NOTES_THESIS.docx#_Toc69911018)

[Figure 3 Athens Metropolitan Area, highlighting the surrounding municipality [27] 20](#_Toc69911019)

[Figure 4 Districts in Municipality of Athens 20](#_Toc69911020)

**List of Tables**

[Table 1 Key Sectors in Smart City and Potential Application [11] 9](#_Toc70437735)

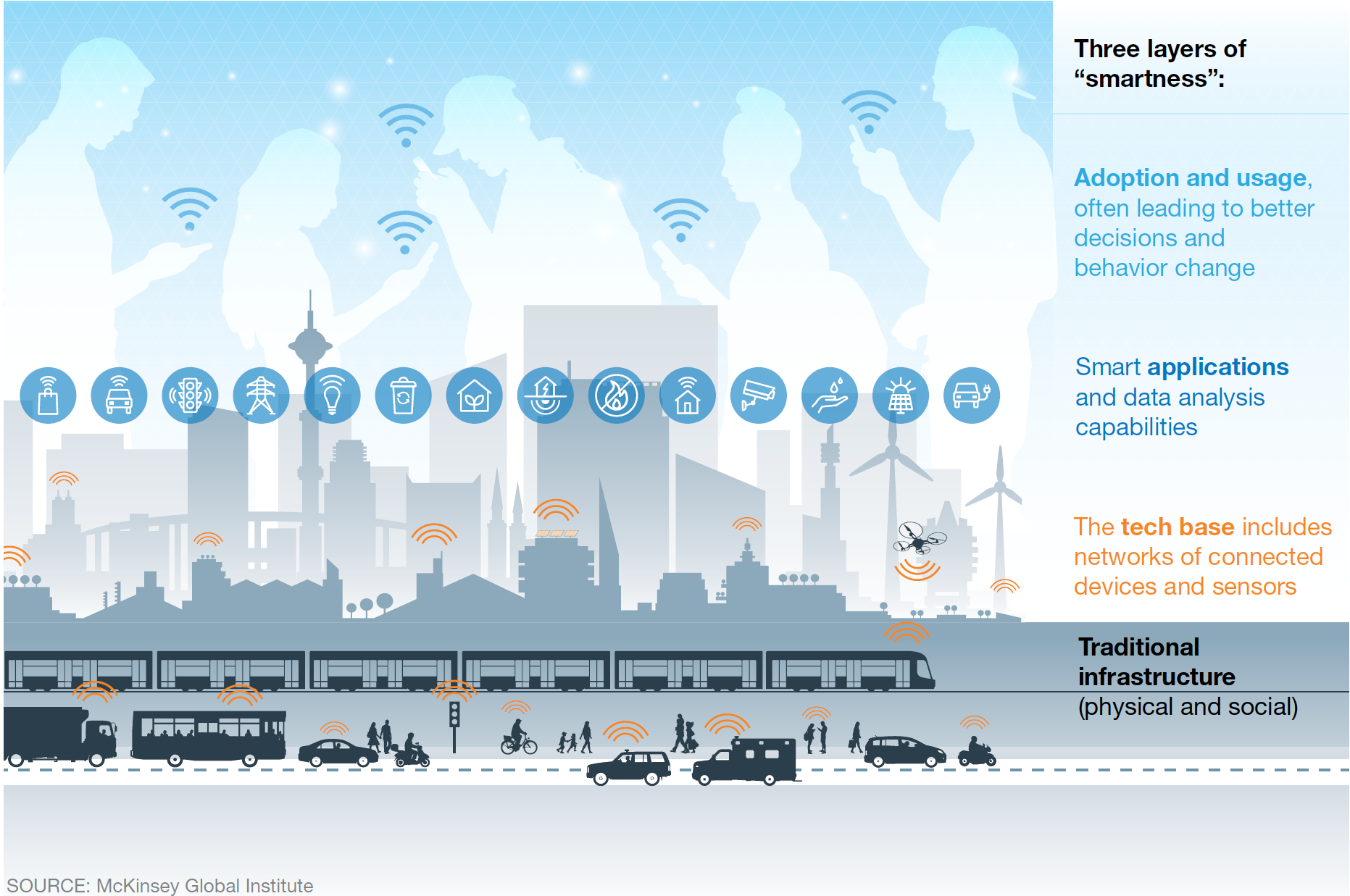
[Table 2: Key Enabling Technologies 9](#_Toc70437736)

# Introduction

The world´s appetite for seamless and smart services has ever been increasing since the technology boom in the end of 20th century. Our daily experiences in cities are increasingly tactile and efficient within the smart intelligent systems: smart devices alert everyday joe of subway delays but assure them that the day’s air quality is good, so he can make an informed decision to cycle to work, staying fit in the process. An entrepreneur applies for a business license and is pleased to find not only a simple digital form with fast approval but ample city data that helps her identify a good location for her new business . A middle-aged woman worried about her aging father living alone is reassured to learn that local healthcare providers can monitor his diabetes and video chat with him in his home. Smart devices are now the new compass to the city, and a smart city is the city ready for future.

Wealth of information at our fingertips navigate us to a more efficient use of our time and resources in our day to day activities. These infinite information feed is continuously fed by layers of sensors and trackers embedded throughout our urban environment: data flows in real time, mined into analytics ecosystems to run hyper-complex city operations and infrastructure often with minimum human interaction. As million more urban dwellers use them to make better decisions, the data accumulations adds up to exponential critical mass of information flow, the city as a whole is more responsive and productive. Less waiting time at transits and queue, more economical solutions to everyday needs. Energy, resources, space, and investment are utilized more efficiently. Ultimately, creating the ecosystem that defines a smart city.

Our growth in technology are also supported by the ever growing urban population, rampant physical development both exacerbated by the digital easiness in doing business and social mobility (up the ladder) is seen to have an adverse effect on how we use our limited resources in cities: land.



**Research Question 1**

**How can data mining impact the real estate industry and what are the current state of art?**

**Research Question 2**

**What data mining techniques can be used to create a prediction model for house prices?**

**This thesis is organized as follow: This section is an introduction on the subject matter. The second section will look into fundamental concept of smart city, big data mining and how they intersect. The scope and methodology of the study is then discussed in section three**

# Smart Cities

Smart city is not relatively new, it has been the topic of discussion since the dot-com boom, and foreground to many science-fiction literature and movies. Data Mining is a key component of smart city infrastructure where big data, considered as a big part of the ecosystem where it feeds and also gain information in a perpetual informational machine. Its definition, etymology and scope however, is ever growing and we look into some relevant literature into the term to make it more understandable in relation to the subject matter of the thesis.

## Smart City Definition

So what makes a city smart? Is it flying cars, sleek urban landscape and isolated high-tech building? Is it its ability to respond to issues in real time? How has our understanding and idea about smart city evolved? We will discussed this further in the next sub-chapter.

An indicative smart city definition comes from ISO/IEC [4] and recognizes the smart and sustainable city as “an innovative city that uses information technology and other means to improve quality of life, efficiency of urban operation and services, and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social, and environmental aspects.” A pretty much common view for most technical urban technologist and professionals alike.

However, the term smart city has been interpreted differently within different context and field of study. Usage of the term “smart” as a branding tool also has been the subject to many debate For example, London- a world ranked smart city, with extensive technology application form wide use of surveillance systems, traffic management and highly robust open data platforms, they are still riddled with unequal growth, empty skyscrapers, homelessness and urban poor [5] where as some of the more interesting use of low technology cities like “sponge cities” in china is solving their urban water management issues with efficient but old school, engineering and planning strategies [6]. However, with a conscious balance of these two approach of low-tech and high-tech strategies can be a sustainable way forward. This gap additionally, shows highlight the potential “smartness” aspect of city building that can be explored and understood.

A comprehensive look by Dameri [7] clearly indicated that there are compulsory elements for a comprehensive a smart city definition: component, boundaries, scope and terminology. They are summarized as “a smart city is a well defined geographical area, in which high technologies such as ICT, logistic, energy production, and so on, cooperate to create benefits for citizens in terms of well being, inclusion and participation, environmental quality, intelligent development; it is governed by a well defined pool of subjects, able to state the rules and policy for the city government and development”. Technology (city management ecosystem and infrastructure), people (socio-economic and socio-culture) and institution (policy, education and governance) then make up as the key stakeholders. We can safely deduce that the technologies in smart city are the key enablers, supported by a robust policy but focuses on outcomes that are beneficial for all its stakeholders.

With that in mind, we can agree that the dimensions and layers can form an ecosystem that comprised all of the elements mentioned by Dameri [7]. Figure 1 shows a ecosystem of smart cities that includes all the stakeholders or as depicted in [8] value creators: cities, utilities, corporations, communities and citizens. They are then supported by “capability layers”: innovation layer, community engagement layer, governance and operational layer, policies and financing layer, data information and marketplace layer, connectivity and security layer and finally the enabling technologies layer. These layers must be synchronized and unified to then deliver the “services” like smart traffic, smart healthcare, smart utility management and smart transportation to name a few. This in the end, cumulatively work in the ecosystem to deliver the desired “outcome” of increased health, socioeconomic and quality of life, seamless and efficient government-city services, ensured sustainability and resiliency in our cities.

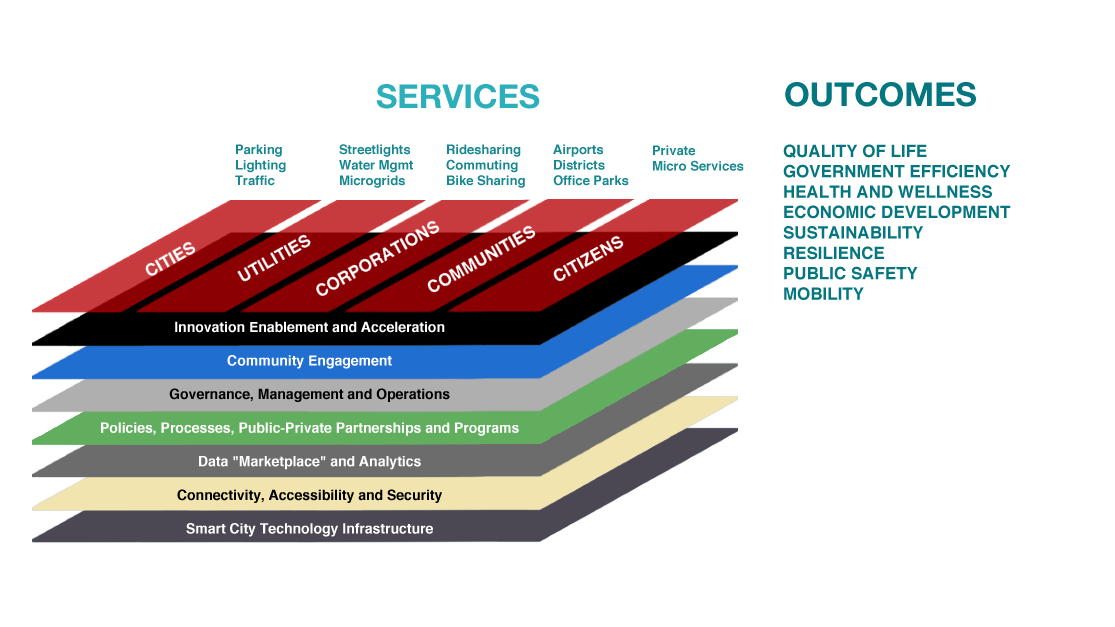


Figure 1 Smart City Ecosystem Framework [8]

#### 2.1 Key challenges in Smart Cities

Smart cities are one of the world’s answer to rapid urbanization. As major cities and government embarks on the digital transformation bandwagon, a number of fundamental challenges hinder the adoption of the smart city ecosystem namely:

* ***Infrastructure***: There is the cost dimension to technologies like sensors and IoT in smart cities to install, maintain and operate them. Additionally,
* ***Reliability of Ecosystem:*** As relianceto the technologies increase, how sure are the system’s reliability in unprecedented times? What would happen if a sub-sector of smart dashboard collapse and what are the strategies to mitigate this.
* ***Privacy and Security:*** A key important factor in adoption of smart cities is the security aspect, both a real and perceptive concern. This includes, issues of personal data privacy, encryption, inconsistent public interface and national security (external threats/hacking etc.) . Regulatory and policy steps should help address this concern, for example the newly minted General Data Protection Regulation (EU) 2016/679 (GDPR) [9].
* ***Citizen Engagement***: Without citizen engagement, the potential of these technologies and adoption of the initiative can be difficult and slow. Public participation and engagement also could reduce the negative perception of technological involvement in their daily life [10].

When these challenges are address in tandem with effective implementation plan, the smart city terminology will be a thing of the past, as growing cities inadvertently will become smart sans the “smart city” characteristics.

## Smart City Infrastructure and Applications

Today, most initiatives or projects entails application of technology associated with smart city are used to ease their daily operations. However, most of them work in silo and are not interconnected [11]. Examples includes transport managements where efficiency and jnrgkoamrñlk is usually solved? Managed without any consideration to other key sectors like healthcare managements, and resource distribution etc. at a cost . vsdknlkdjnsdb.

Future of a smart city data platform or information system where real time, historical, descriptive and predictive knowledge from big data can inform daily operations, urban growth, city ecosystem: to achieve the desired output. There are myriads of new terminology to describe these user interface both for city governance and the public. Smart City Geospatial Dashboard is among the popular term in smart city governance, where it can gather, visualize, analyze and advise on real-time situation, operation and affect future planning [12]. The other is Digital Twin Cities, a simulation process that makes full use of physical models, sensors, historical data of operation, etc. to integrate information of multi-discipline, multi-physical quantities, multi-scale, and multi-probability that takes in a virtual space[13]. A Hollywood portrayal of digital twin is not uncommon, in the movie Elysium, a simulated 3D holographic cityscape of the planet, complete with infographics that portray real time city operations and future projections consume the major space in the planets command centre. This however, is out from the scope of this study, but important to note as a leap into that direction.

A seamless smart city infrastructure would require minimum human intervention, but it will affect maximum human quality of life in the city. They are key sectors that the smartness can be applied as depicted in Table 1.

|  |  |  |
| --- | --- | --- |
| Key Sectors | Application | Projects |
| Transport and  Traffic Management | Vehicular Congestion  Frequency of usage  Accidents  Commuting Optimization Features  Traffic data processing  Signal management | City Brain [14]  Digital Twin Cities [13]  Smart Light  Electric Vehicle Network |
| Population  Distribution  Mapping | Land Use and Urban Planning  Congestion mapping  Demographic distribution  Population mapping  Gender distribution | Super City Planning [15]  Smart Real Estate [16] |
| Utilities Distribution  Network | Resource availability  Resource reliability  Utilities infrastructure  Utilities usage | Smart Grid [17]  Smart Building [17]  Smart Meter |
| Health Care  Management | Frequency of visit  Type of treatments  Type of incidents  Emergency response time | City Brain [14]  Smart Health  Seat Pleasant [18] |
| Disaster Mitigation  and Security | Historical data  Simulation study  Security assessment  CCTV  Surveillance data | Environmental Monitoring |

Table 1 Key Sectors in Smart City and Potential Application [11]

Smart city ecosystem and application will only grow in time, as exploration into disruption of technologies into every aspect of city life. City Brain for example is an initiative lead by Ali Baba Cloud Technologies, that enable city managers and government in cognizing, transforming, and operating cities [14]. City brain enables utilization of massive data in real time, identity trends and pattern through machine learning and formulate strategic solution based on global resources and dynamics (surpass local level awareness).

Key enabling technologies in this sectors can be

|  |  |
| --- | --- |
| Key Technologies | Products |
| Internet of Things | Sensors and Data hubs |
| Smart Dust |  |
| Smart Phones |  |
| Cloud Computing |  |
| Big Data and Open Data |  |
| Smart Grid |  |
| Smart Home Devices | Nest[19] Google Home [20] |

Table 2: Key Enabling Technologies

* + - Enabling technology :
    - What are the IOt, sensors, what are the level of adoption?

Moving forward, we shall discuss a major component of the smart city that pertains to data architecture which is data mining.

## Smart Cities in Context

Adoption around the world in pursuit of the smartest cities tittle has resulted in some impressive initiative and project of multitude of scale and size. Investment into technology and infrastructure has seen some cities dominating the rankings year on year. These rankings are considered an effective instruments for most cities to attract investments and competitive edge to their cities [21]. This ranking tool is also beneficial for cities to asses their own strength and weakness while also acting as an external key performance indicators (KPIs) to their smart city initiatives. There are many city rankings available publicly however, its reliability as an assessment tool has also been in question. Academics and professionals has also been critical of city rankings, where increased competition between city can threaten long term development growth which might be a negative consequences of cutting corners (deregulation, structural and spatial problems and risk delicate socio-economic balance)[21]. It can be a double edged sword, but a holistic and well-documented, methodically advanced rankings conducted by universities or economic research institutes with a clear focus on context (country level or continent) can be a good reference. Cities then are able to position themselves in the region or continent and focus on improving sectors that are lacking in their profile.

Cities in Motion Index (CIMI) is widely considered an exhaustive ranking of smart cities that include 101 indicators across 9 key dimensions: human capital, social cohesion, the economy, governance, the environment, mobility and transportation, urban planning, international projection, and technology, reflecting both objective and subjective data to offer a comprehensive view of each city [22]. London and New York a big and populous city has been dominating the ranks for the 6 consecutive years, with notable presence of some smaller city-state like Singapore and Hong Kong in their top 10. Interestingly, there is an artificial line drawn with each city’s approach where some cities like Singapore, London, New York has opted for a high-tech centric and significant infrastructure investment approach and the other side of community-centric approach in cities like Copenhagen, Amsterdam and Vienna.

Athens at 96th ranking for example has taken the latter approach, where it has lined it strategies in low-hanging, short term projects, where effective medium-cost smart city projects focuses on collaboration between existing technology, communication and energy companies with academics and communities to execute them. This is reflected in the CIMI where some of the low ranked indicators are social cohesion (148th), governance (148th), urban planning (142nd) and economy (109th) while doing well on human capital (71st) , environment (57th) and international projection (52nd). Athen’s 12-month Digital Roadmap focuses on five key elements: infrastructure (networks and internet access), public spaces and residents, government digital services, citizen engagement and innovation [23]. In this dissertation, we will look further on Athens data architecture and infrastructure in the following chapter in relation to the dissertation problem statement.

//Smart city application (benefit to stakeholders)

//IoT ( what is it) how data mining in essential? What is the current data mining field progress (5V)

//Why smart city is important, to remind is to improve quality of life ( set parameters that can determine a smart city. Not necessarily technology centric but it has to start with people. Sometime we forget that we work so hard to make the city smart in terms of technology but it also increases disparity and inequality in cities, how to address that\_?

//Technology should be used as a tool to then create the solution , it’s the means for a smart city but not the end result.

# Big Data, Data Mining in Smart Cities

*Big data is an essential component of a smart city. Based on Figure 1, data mining would be characterized in the Data “Marketplace and Analytics layer, where big volumes of data from interconnected Internet of Things (IoT) and sensors that gather, store, share and communicate data, with an aim to improve efficiencies across the smart city infrastructure. Big data is a foundational tool to design efficient and livable cities for people.*

## Big Data in Smart Cities

Big data is an essential component of a smart city. Based on Figure 1, data mining would be characterized in the Data “Marketplace and Analytics layer, where big volumes of data from interconnected Internet of Things (IoT) and sensors that gather, store, share and communicate data, with an aim to improve efficiencies across the smart city infrastructure. Big data is a foundational tool to design efficient and livable cities for people.

### General Concepts

Knowledge discovery of a vast amount of data can be overwhelming for an average human to comprehend. To make sense of all the data coming from variety of sources: geospatial data, traffic data, vehicular traffic data, crime statistics, etc. we need big data analytics. Through algorithms, city stakeholders can draw connections across distinct sources to reveal useful information. Figure 2 shows a popular keywords relating to big data obtained from the research done by [24]. This shows the multidisciplinary use of big data in the world today. Its impact has been enormous in propelling efficient and robust solution to most industries, and the movement will only grow from this point on.



Figure 2: Static tag cloud visualization of key terms appearing in abstracts of Big Data-related papers [23]

Big data has evolved rapidly and vastly, the sheer nature of the discipline has allowed various definition of big data to specific industry and no universally accepted definition has existed. However, authors in [24]–[27] has defined big data in the followings 5Vs :

* ***Variety:*** This refers to complexity of big data and also its ability to integrate structured and unstructured data (whether it is text, images, voice, videos, streaming data, signals or other types of data) from multiple sources into a comprehensive resource or database.
* ***Volume:*** This relates to the sheer nature of big data that will challenge existing hardware, software in terms of storage, processing, analysis and visualization capabilities. This also inevitable mean the exponential growth will affect the capability of big data to generate knowledge and insights.
* ***Velocity:*** Thisrefers to the speed with which the data is generated, analyzed and reprocessed into the platforms. Today this is mostly possible within a fraction of a second, known as real time. Velocity also raises a new concern on data ageing, in question of the data validity [28]. For example, time sensitive data like traffic and crime surveillance.
* ***Veracity (Validity):*** This refers to data quality and authenticity that ensure the credibility of the data. A bad data can be a liability and a burden to any decision based on the data analytics output.
* ***Value:*** This refers to the potential added value to the user of big data (business, government, planners etc)

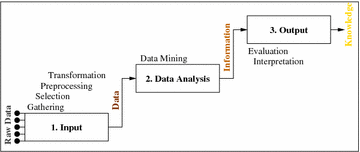
Interestingly, some researchers argue that the term “big” in big data will fade over time and “data” will be self describing thus naturally include all the big data characteristic mentioned above [28]. The impact of big data in smart city is unmeasurable, where the success of a smart city, depends on how robust the use of big data in its infrastructure.

### Big Data Value Chain Operators:

Big data roles min the new technology age is undisputable and businesses, government and people are incorporating data architecture into their organization and this is only the beginning. In smart city specifically, big data innate characteristic of volume, velocity and variety poses a challenge to the smart city framework. Therefore, an approach by author in [29] is the Knowledge Discovery in Database (KDD) model is used to streamline the big data analytics. In the model (Figure 3), three operators are responsible : input, data analysis and output. The value chain operators are made up of subsystems and the information flow in KDD is described as a string of actions needed to generate value and insight from big data. The operators includes these key processes:

* ***Input:*** This is a critical phase where data are extracted (gathered), selected, pre-processed and transformed to useable data that are cleaned and aligned with the 5Vs attribute. The selection operators are tasked to integrate and select relevant information from the gathered database. The preprocessing operator then, would filter, clean and detect the inconsistent, missing or incomplete data. Then the data would be transformed into the variety of data formats into data-mining-capable format. Data Input phase effectively reduce the complexity and scale the data that is appropriate for data analysis.
* ***Data Analysis:*** This phase is responsible for discovering the hidden patterns or rules from the datasets, most investigators in data mining field use the term to describe how they hone the “ground” (i.e. raw data) into “gold nugget” (i.e. information or knowledge)[29]. Data mining techniques and tool will be discuss further in the next sub-chapter.
* ***Output:*** This covers the data-driven business activities that need access to data, its analysis, and the tools needed to integrate the data analysis within the business activity. The mined data is evaluated for its accuracy and veracity. Finally the data usage goes through interpretation where organization use the insight to make informed decision on strategies or measure existing performance criteria etc.

Figure 3: Knowledge Discovery in Databases model [28]



### Platforms

To address the challenging nature of big data, platform scalability is key in ensuring a successful work flow. A big data platform can scale vertically and horizontally which is also known as scale up or out [28]. Vertical scaling implies additional computing power like RAM and CPUs while working on a single Operating System (OS). In the opposite, horizontal scaling implies a division of datasets over multiple parallel servers or *shards*. So more machines are added as much as needed to improve the platform’s performance. Both has pros and cons, which ultimately depends on the situation at hand. A vertical scaling is great in handling due to a single operating system, but it can be costly in comparison. While horizontal scaling provides you the ability to process your data in smaller chunks and possibly with less time but, managing multiple instances of operating system can be complex. Sample of vertical scaling are Graphics Processing Unit (GPU), High-Performance Computing Clusters (HPC), and Multicore processors, and Field Programmable Gate Arrays (FPGA) and some of the popular horizontal scaling platforms are Apache Hadoop and Map Reduce.

However, most research or projects in smart cities favors horizontal scaling due to the multi domain nature of smart city that is ever expanding thus its only sensible to rely on a horizontal scale out approach, as the smart city services or infrastructure grow.

## Data Mining

Big data is a conceptual term that describe the large amounts of data whereas data mining refers to a technique to analyze data. In Figure 3, Data mining belongs to the data analysis phase of the knowledge discovery or KDD. Techniques in data mining is not as seldom used in urban analytics as Geographic Information System (GIS) but it presents an opportunity for analyst or researcher to combine these two techniques to get detailed differentiation in the urban forms.

### Data Mining functionalities.

In a world of endless streams of data, there is a great need to transform them into something of value or knowledge. It is established that data mining is an essential step in knowledge discovery process (Figure 3). Data mining is a process of knowledge and pattern discovery from large amount of data (big data). Data can be sourced from various database, data warehouse, the web or real time data that are streamed directly into the system.

Data mining can be used to run a predictive or descriptive task aptly named to their functionalities. Predictive mining task performs induction on existing data to make prediction and descriptive mining task characterize present properties of the data in the datasets [30]. There are a number of data mining functionalities that are used to specify the type of patterns to be retrieved from data mining tasks. The data mining functionalities is as follows:

* ***Class and Concept description:*** All datasets can be associated with classes and concept that has a description that describes the individual classes and concepts in concise and yet precise term. These descriptions can be derived through data characterization and data discrimination.
* ***Association Analysis:*** It entails the discovery of association rule or frequent patterns in data. Mining frequent patterns precedes to the discovery of fascinating associations and correlations within data [30]. However, the rule can be discarded or deemed uninteresting if it doesn’t satisfy a minimum *support* threshold and a minimum *confidence* threshold. *Support* is an indication of frequency of the itemset appearing in the dataset, where *confidence* means is how frequently is the rule is proven to be true. Example of use : Mining frequent items bought together in grocery purchases, then making decision to shelf them next to each other.
* ***Classification:*** It consist of a process of acquiring a model (function) that describes or differentiates the data classes or concepts. Classification model obtained from the analysis into a *training data* ( i.e. data objects for which the class labels are known). The model is then used to predict the class label of objects when it is unknown. Example of use : Categorizing applicant for a credit card in to “good”, “bad” credit rating by analyzing the attributes through select techniques.
* ***Cluster Analysis:*** It is a process which objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity [30]. That is, clusters are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters based on distance.
* ***Outlier Analysis:*** Its is also known as anomaly detection where datasets that may not conform to the normal or general behavior of the data. Outliers might be spotted using statistical tests that presume a distribution or probability model for the data, or using distance calculations where objects that are remote from any other cluster are considered *outliers*. Example of use: It can be used to detect fraudulent usage of credit card or any other illegal activities.

A data mining system has the potential to discover millions of pattern and rules based on large amount of data that is available this day and age. With that also, raises questions about their interestingness or relevance of said patterns to the organization’s benefits etc. There are measure to hone these discovered pattern into ranks, filtering the uninteresting ones, for example through the *objective measures of pattern interestingness or subjective interestingness measure.* Measures of pattern interestingness are vital for the efficient discovery of patterns by target users by pruning the patterned discovered that do not satisfy a prespecified interestingness parameters.

### Evaluation and Metrics in Data Mining

### Challenges in Data Mining

Data mining as a field is still at its adolescent stage where it will be ever-growing, and ever-expanding. To continue in this growth uptrend, a few key challenges needs to be addressed to further improve and enrich the field. They are articulated as follows:

* ***Security and Social Challenges:*** Despite the immense benefit of knowledge discovery in data mining, there is a real concern in breach of personal privacy and security and increasingly negative public perception to the intrusive nature of the technology [31]. The undesired discovery of patterns and access to possible sensitive data can further hinder the adoption of data mining in smart cities. However, a sub-field of study privacy preserving data mining (PPDM) is gaining traction in its development recently, as to ensure safeguards of sensitive information will still maintaining potential utility of information . This is a welcomed initiative to improve real privacy and safety concern while ensuring confidence to the public of their personal privacy and security.
* ***User Interface:*** The knowledge discovery process is only beneficial if it is noteworthy and comprehensible by the user. There are important consideration for data mining platforms allows flexible, creative and interactive interface to allow dynamic exploration of the focus of the task such as possible by-products of the initial query (i.e. unexpected pattern discoveries). The interface should also consider specific user background knowledge or previous discovered patterns.
* ***Data quality and management:*** The interoperability nature of smart city data structure means the data needs to adhere to a certain level of quality. Quality of Information (QoI) of data from multi sources of various smart city infrastructure Internet of Things (IoT) such as weather, traffic, disaster and pollution can be different and complex [31]. Of course, the data typology and quality is dependent on the functional requirement as well, thus a holistic understanding of the application and use of these data within the smart city framework can help clarify the minimum level of data standardization and quality, thus enabling data integration and aggregation for a high level knowledge discovery. (means data is filtered for relevance for only high level understanding). Data often contains noise, missing value and uncertainty, thus proper preprocessing and cleaning measures is important to ensure data veracity.
* ***Complex Data***: Data can come from variety of sources ands some organization use different data structures. Diverse sources generate different type of data like structured and unstructured can prove very challenging to streamline and prepped for the system. The creation of effective and efficient data mining tools for varied applications remains a difficult and dynamic area of research.
* ***Performance:*** the data mining system performance depends on the efficiency and scalability of the algorithm. A faulty or inappropriate algorithm to the specific task can affect the overall performance of the system.

# Real Estate & Smart City

Smart city initiatives are expected to increase the quality of life, efficiency in city operations and productivities for cities in the long run. It is forecasted that smart city projects will reinforce the global cities to new heights, and emerging cities to be competitive as technology become more accessible and cheap. Some cities are doubling down with big investment into smart cities while some cities are doing smaller scale interventions. Nevertheless, we will see increased uptake in technology and its byproduct: big data

## Impact of Smart City in Real Estate

Most cities embarking on a smart city initiative are banking on the notion of improved efficiency, productivity and overall quality of life for its citizens. The trickle down ideals of job creation thorough expansion of the physical, information and communication infrastructure are the reason smart city initiative is viewed the next big thing, in line with industrial revolution 4.0. With allocation of special economic zones and policies that supports smart cities, a dynamic synergy between public and private entities will pull capital and investment towards the initiative and create a robust environment for smart city implementations [32]. The development of smart city will of course benefit the real estate industry and increase demand for housing and other asset class such as hotels, retails and offices. They are intrinsically linked.

These interest and attention to the real estate aspect of smart city has give birth to a new concept of smart real estate [16].

Below are a few impacts of smart city implementation to real estate:

* Encourage public-private partnership in smart real estate projects
* Better quality of real estate product
* Equitable utility and service distribution in new real estate development
* Ability to predict real estate trends and volatility
* Ensure a stable supply and demand of housing ensuring affordability
* Create a healthy demand for smart management and other smart features
* Inform policy maker on impact to the land use and real estate policy in cities

In mature cities most development in smart city will be mostly refurbishment and retrofit, but in new and emerging cities, smart city development are often from scratch as an answer to rising population [33]. In this case, the impact to real estate would vary where mature cities often have more complex variables that affect the real estate supply, price and overall trend. For example, in newer smart cities, housing affordability can be a controlled and planned to ensure a well distributed land use while it will be difficult to do so in mature cities. However, in mature cities, government can identify opportunity places and site for housing, amenities, public spaces opportunities by using sophisticated algorithm that make use of big data. Mature cities can adopt a retrofit strategy…….

* Smart home to smart building, smart meter, smart grid
* Big data availability as a result of smart city implementation.

## Data mining in Real Estate

Real estate growth in major cities are exacerbated mostly by a perceived demand and supply where traditional market analysis uses a comparative market analysis (CMA) tool that is low in granularity that only include a small number of metrices. With availability of big data, a bigger metrics can be included in the process in analyzing the real estate market. An efficient way to do this analysis of big data is through data mining.

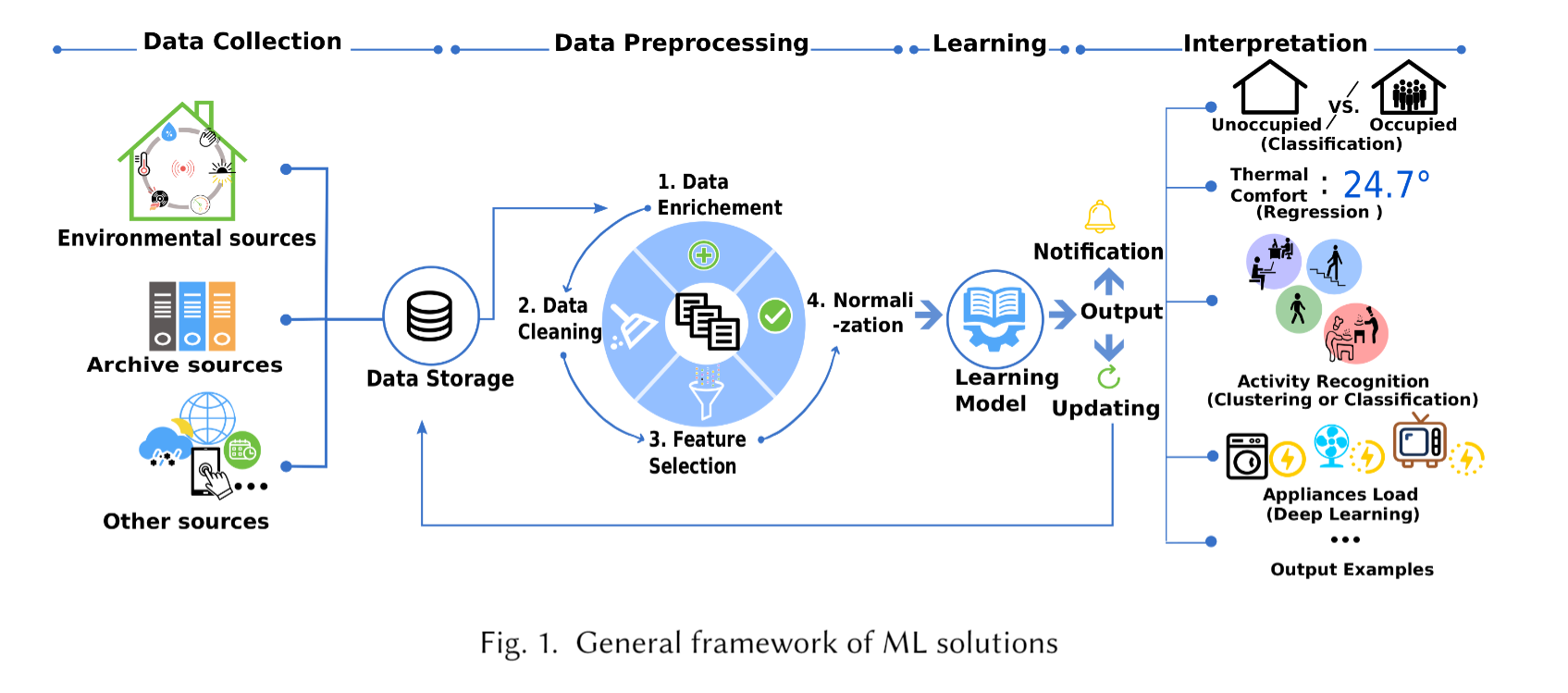


Figure 4: Data mining and machine learning framework of data flow

There is a wide range of application of data mining or machine learning within the real estate field. Both government and private sector can benefit from the knowledge gain from big data, for example, authorities can track development progress, expedite permit planning process and ensure public policy interests are protected while encouraging a dynamic and competitive commercial real estate while private developers can identify opportunities and attractive projects to work on. These technology can be used to forecast supply, market trends, valuation and client management in commercial real estate [34]. Even other private sectors like banking can use forecast data to strategize capital requirement, possible interest revenue and losses.

There are various level of data mining and machine learning application in the field. Some are used at proxy level, where usage are expanded into customer service and consumer apps while more sophisticated application of the technology can further into analytics in building automation system (smart buildings and homes), automation of property management and analysis of the real estate market [35]. Table 3 shows the potential application of data mining and machine learning in real estate and some promising companies that are currently pursuing them.

|  |  |  |
| --- | --- | --- |
| Application Area | Description | Companies |
| Building Automation Systems | Analytical framework that includes Internet of Things (IoT) (i.e. sensors, camera and etc) for electrical, lighting, security and transportation system in the building. Data collected also can support the smart city database and allow predictive analytics for other smart city services [36]. | Pointgrab [37]  BuildingIQ [38] |
| Property Management | Automation of property management job function like managing and maintaining assets in big real estate portfolio [39]. It can predict maintenance requirement of building services, rental payment, lease extensions and other concierge services like tenant reports etc. | VTS [40]  AppFolio [41]  Zen Place [42] |
| Market Analysis | Valuation process and understanding the real market analysis is achieved by machine learning. It can help investors understand the market better and make informed decision with a bigger variables and with highly specific needs. | Zillow [43] |

Table 3: Application of data mining and machine learning application in real estate

This paper will focus on the market analysis application of data mining and machine learning, where a big part of the real estate industry is dictated by the pricing mechanism. Various literature has attempted to create a prediction model for real estate prices and also a forecasting model for future sales and supply of assets into the market.

## Selected Real Estate Market

London was selected as a study area based on the following reasons: London is the world premier smart city and have been consistently ranked as a top smart city [22], Greater London Authority (GLA) has a robust blueprint and smart city plan [44] and London has a robust open data platform in London Data Store [45]. London´s plan sets out to transform London as a leading smart city in the world, which is articulated in their five (5) mission statement:

* **Mission 1:** More user-designed service
* **Mission 2:** Strike a new deal for city data
* **Mission 3:** World-class connectivity and smarter streets
* **Mission 4:** Enhance digital leadership and skills
* **Mission 5:** Improve city-wide collaboration

# Methodology

The aim of this study is to new technology to understand the real estate market. Our main strategy is to develop a data mining / machine learning techniques to forecast housing price. This forecast model can be used to identify key opportunities areas for real estate investment, regeneration policy and housing strategies. This would help policymakers make long term strategies, private citizens making financial plans for personal real estate investment and private sector to use the model to make decision on capital investment and business strategies accordingly.

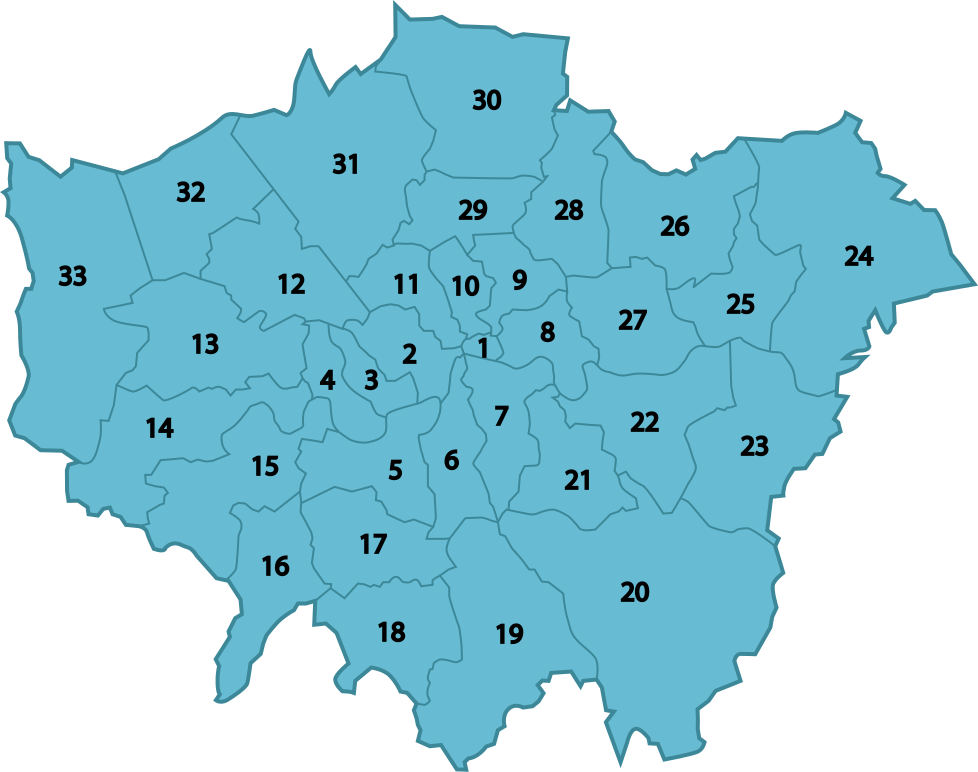
The idea was to extract data from open data platform, transform data from various sources into a readable format using several algorithm and techniques for data pre-processing, mining and then data visualization. For this project, three main tasks is aligned to achieve the desired output:

1. Data Gathering and Selection
2. Prelim EDA
3. Data Pre-processing and Transformation
4. Model Selection for Prediction

## Data Extraction

The datasets has been extracted from London Data Store. It is released under UK Open Government License v2 and v3. This is an open data platform run and maintained by the Intelligence Unit of the Greater London Authority (GLA). This dataset contained attributes such as price index, density and size of dwellings, sales records, crime records and population density per borough in London. Another dataset that define the borough shapefiles are also extracted from London Data Store is used to create visualization.

|  |  |
| --- | --- |
| 1. [City of London](https://en.wikipedia.org/wiki/City_of_London) 2. [City of Westminster](https://en.wikipedia.org/wiki/City_of_Westminster) 3. [Kensington and Chelsea](https://en.wikipedia.org/wiki/Royal_Borough_of_Kensington_and_Chelsea) 4. [Hammersmith and Fulham](https://en.wikipedia.org/wiki/London_Borough_of_Hammersmith_and_Fulham) 5. [Wandsworth](https://en.wikipedia.org/wiki/London_Borough_of_Wandsworth) 6. [Lambeth](https://en.wikipedia.org/wiki/London_Borough_of_Lambeth) 7. [Southwark](https://en.wikipedia.org/wiki/London_Borough_of_Southwark) 8. [Tower Hamlets](https://en.wikipedia.org/wiki/London_Borough_of_Tower_Hamlets) 9. [Hackney](https://en.wikipedia.org/wiki/London_Borough_of_Hackney) 10. [Islington](https://en.wikipedia.org/wiki/London_Borough_of_Islington) 11. [Camden](https://en.wikipedia.org/wiki/London_Borough_of_Camden) 12. [Brent](https://en.wikipedia.org/wiki/London_Borough_of_Brent) 13. [Ealing](https://en.wikipedia.org/wiki/London_Borough_of_Ealing) 14. [Hounslow](https://en.wikipedia.org/wiki/London_Borough_of_Hounslow) 15. [Richmond upon Thames](https://en.wikipedia.org/wiki/London_Borough_of_Richmond_upon_Thames) 16. [Kingston upon Thames](https://en.wikipedia.org/wiki/Royal_Borough_of_Kingston_upon_Thames) 17. [Merton](https://en.wikipedia.org/wiki/London_Borough_of_Merton) 18. [Sutton](https://en.wikipedia.org/wiki/London_Borough_of_Sutton) 19. [Croydon](https://en.wikipedia.org/wiki/London_Borough_of_Croydon) 20. [Bromley](https://en.wikipedia.org/wiki/London_Borough_of_Bromley) 21. [Lewisham](https://en.wikipedia.org/wiki/London_Borough_of_Lewisham) 22. [Greenwich](https://en.wikipedia.org/wiki/Royal_Borough_of_Greenwich) 23. [Bexley](https://en.wikipedia.org/wiki/London_Borough_of_Bexley) 24. [Havering](https://en.wikipedia.org/wiki/London_Borough_of_Havering) 25. [Barking and Dagenham](https://en.wikipedia.org/wiki/London_Borough_of_Barking_and_Dagenham) 26. [Redbridge](https://en.wikipedia.org/wiki/London_Borough_of_Redbridge) 27. [Newham](https://en.wikipedia.org/wiki/London_Borough_of_Newham) 28. [Waltham Forest](https://en.wikipedia.org/wiki/London_Borough_of_Waltham_Forest) 29. [Haringey](https://en.wikipedia.org/wiki/London_Borough_of_Haringey) 30. [Enfield](https://en.wikipedia.org/wiki/London_Borough_of_Enfield) 31. [Barnet](https://en.wikipedia.org/wiki/London_Borough_of_Barnet) 32. [Harrow](https://en.wikipedia.org/wiki/London_Borough_of_Harrow) 33. [Hillingdon](https://en.wikipedia.org/wiki/London_Borough_of_Hillingdon) |  |



The first dataset initially contained 13549 instances and 7 attributes. This dataset is a monthly variable of each housing condition in all 33 borough from the year 1995 to 2020.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| date | Timestamp | Time period of the record |
| area | Varchar | Name of the borough |
| average\_price | Number | Mean house price |
| code | Number | Area code according to Authority |
| houses\_sold | Number | Number of houses sold |
| no\_of\_crime | Number | Number of crime committed |
| borough\_flag | Number | Indication of borough belonging to London Authority jurisdiction |

The second dataset is a yearly variable that contained 1071 instances and 12 attributes. This dataset is recorded from the year 1999 to 2019.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| code | Number | Area code according to Authority |
| area | Varchar | Name of the borough |
| date | Timestamp | Time period of the record |
| median\_salary | Float | Median salary of the individuals living in the area |
| life\_satisfaction | Float | Life satisfaction of the individuals living in the area |
| mean\_salary | Float | Mean salary of the individuals living in the area |
| recycling\_pct | Float | Percentage of households recycling |
| population\_size | number | Number of people living in the area |
| area\_size | float | Size of the area (in hectares) |
| borough\_flag | Number | Indication of borough belonging to London Authority jurisdiction |

## Data Preprocessing and Transformation

The data pre-processing phase entails the process of cleaning and transforming the data that’s relevant to the output of the mining process. A few attributes like dates and year needs to be separated and data from yearly datasets need to be combined into the monthly datasets.

An initial Exploratory Data Analysis was done to get a look into the valuable information of the datasets and number of missing values or if there is any unimportant attributes.

1. Drop redundant column
2. Drop any rows or column with more than 50% missing value
3. Merge yearly and month dataset

## Model Selection

1. COMPARE **ARIMA, RANDOM WALK, SARIMA VS VARMA**
2. Will test the accuracy model based on RSME AND R2 of each model and select the bext model
3. Train\_test\_split 0.8 :0.2
4. I will use Microsoft azure to do find the best algorithm or model evaluation and then will test the selected model for

# Conclusion and future work

Bibliography

[1] G. Chi, Y. Liu, Z. Wu, and H. Wu, ‘Ghost Cities Analysis Based on Positioning Data in China’, *ArXiv151008505 Phys.*, Nov. 2015, Accessed: Mar. 05, 2021. [Online]. Available: http://arxiv.org/abs/1510.08505

[2] J. J. V. Ogáyar, ‘The formation of marginal urban spaces in cities in globalization’, *Iguales y diferentes*. https://identification.hypotheses.org/1447 (accessed Apr. 05, 2021).

[3] J. Jacobs, ‘The Death and Life of Great American Cities’, 2016, pp. 94–109. doi: 10.1002/9781119084679.ch4.

[4] ‘ISO/TS 37151:2015(en), Smart community infrastructures — Principles and requirements for performance metrics’. https://www.iso.org/obp/ui/#iso:std:iso:ts:37151:ed-1:v1:en (accessed Mar. 09, 2021).

[5] B. Sterling, ‘Stop Saying “Smart Cities”’, *The Atlantic*, Feb. 12, 2018. https://www.theatlantic.com/technology/archive/2018/02/stupid-cities/553052/ (accessed Apr. 12, 2021).

[6] F. K. S. Chan *et al.*, ‘“Sponge City” in China—A breakthrough of planning and flood risk management in the urban context’, *Land Use Policy*, vol. 76, pp. 772–778, Jul. 2018, doi: 10.1016/j.landusepol.2018.03.005.

[7] R. Dameri, ‘Searching for Smart City definition: a comprehensive proposal’, *Int. J. Comput. Technol.*, vol. 11, p. 2544, Oct. 2013, doi: 10.24297/ijct.v11i5.1142.

[8] ‘A Smart City Ecosystem Framework for building sustainable smart cities’, *Strategy of Things*, Jan. 24, 2018. https://strategyofthings.io/smart-city-ecosystem (accessed Apr. 17, 2021).

[9] ‘EUR-Lex - 32016R0679 - EN - EUR-Lex’. https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed Apr. 17, 2021).

[10] E. Ismagilova, L. Hughes, N. P. Rana, and Y. K. Dwivedi, ‘Security, Privacy and Risks Within Smart Cities: Literature Review and Development of a Smart City Interaction Framework’, *Inf. Syst. Front.*, Jul. 2020, doi: 10.1007/s10796-020-10044-1.

[11] K. Hein and F. B. A. Rosman, ‘Data-driven Approach for Smart Urban Planning Tool Development’, p. 8.

[12] C. Jing, M. Du, S. Li, and S. Liu, ‘Geospatial Dashboards for Monitoring Smart City Performance’, *Sustainability*, vol. 11, no. 20, Art. no. 20, Jan. 2019, doi: 10.3390/su11205648.

[13] L. Deren, Y. Wenbo, and S. Zhenfeng, ‘Smart city based on digital twins’, *Comput. Urban Sci.*, vol. 1, no. 1, p. 4, Mar. 2021, doi: 10.1007/s43762-021-00005-y.

[14] X. Hua *et al.*, ‘The City Brain: Practice of Large-Scale Artificial Intelligence in the Real World’, *IET Smart Cities*, vol. 1, May 2019, doi: 10.1049/iet-smc.2019.0034.

[15] M. M. Rathore, A. Paul, A. Ahmad, and G. Jeon, ‘From Smart City towards Next Generation Super City Planning’, *Int. J. Semantic Web Inf. Syst.*, vol. 13, no. 1, p. 20, 2017.

[16] E. Allameh, M. Heidari, de Vries, H. J. P. Timmermans, and J. Beetz, *Smart Home as a smart real estate, A state of the art review*. 2011.

[17] D. Al-Abri, A. Malik, S. Al-Saadi, M. Albadi, Y. Charabi, and N. Hosseinzadeh, ‘Smart Grids and Smart Buildings’, 2021. doi: 10.1007/978-1-4614-6431-0\_78-2.

[18] mHealthIntelligence, ‘MD City Integrates Telehealth With Smart City Hub to Monitor Coronavirus’, *mHealthIntelligence*, Mar. 23, 2020. https://mhealthintelligence.com/news/md-city-integrates-telehealth-with-smart-city-hub-to-monitor-coronavirus (accessed Apr. 21, 2021).

[19] ‘Real Savings | Nest’. https://nest.com/thermostats/real-savings/ (accessed Apr. 21, 2021).

[20] ‘Google Nest (smart speakers)’, *Wikipedia*. Apr. 19, 2021. Accessed: Apr. 21, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Google\_Nest\_(smart\_speakers)&oldid=1018706585

[21] R. Giffinger, H. Gudrun, Gudrun, and G. Haindlmaier, ‘Smart cities ranking: An effective instrument for the positioning of the cities’, *ACE Archit. City Environ.*, vol. 4, Feb. 2010.

[22] ‘IESE Cities in Motion Index 2020 | Cities in Motion’. https://blog.iese.edu/cities-challenges-and-management/2020/10/27/iese-cities-in-motion-index-2020/ (accessed Apr. 17, 2021).

[23] ‘Athens shows how smart cities can do more with less’, *Smart Cities World*. https://www.smartcitiesworld.net/special-reports/special-reports/athens-shows-how-smart-cities-can-do-more-with-less (accessed Apr. 17, 2021).

[24] A. De Mauro, M. Greco, and M. Grimaldi, ‘What is big data? A consensual definition and a review of key research topics’, *AIP Conf. Proc.*, vol. 1644, no. 1, pp. 97–104, Feb. 2015, doi: 10.1063/1.4907823.

[25] D. Agrawal *et al.*, ‘Mind your Ps and Vs: A perspective on the challenges of big data management and privacy concerns’, in *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, Feb. 2015, pp. 1–6. doi: 10.1109/35021BIGCOMP.2015.7072814.

[26] Z. Allam and Z. A. Dhunny, ‘On big data, artificial intelligence and smart cities’, *Cities*, vol. 89, pp. 80–91, Jun. 2019, doi: 10.1016/j.cities.2019.01.032.

[27] ‘Big Data Impacts Data Management: The 5 Vs of Big Data’. https://davebeulke.com/big-data-impacts-data-management-the-five-vs-of-big-data/ (accessed Apr. 26, 2021).

[28] A. M. S. Osman, ‘A novel big data analytics framework for smart cities’, *Future Gener. Comput. Syst.*, vol. 91, pp. 620–633, Feb. 2019, doi: 10.1016/j.future.2018.06.046.

[29] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, ‘Big data analytics: a survey’, *J. Big Data*, vol. 2, no. 1, p. 21, Oct. 2015, doi: 10.1186/s40537-015-0030-3.

[30] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. 2000.

[31] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, ‘Information Security in Big Data: Privacy and Data Mining’, *IEEE Access*, vol. 2, pp. 1–28, Jan. 2014, doi: 10.1109/ACCESS.2014.2362522.

[32] P. Lecomte, ‘New boundaries Conceptual framework for the analysis of commercial real estate in smart cities’, *J. Prop. Invest. Finance*, vol. 37, pp. 118–135, Jan. 2019, doi: 10.1108/JPIF-10-2018-0083.

[33] ‘Are smart cities the answer to rising populations?’, *the Guardian*, Jan. 28, 2013. http://www.theguardian.com/sustainable-business/blog/smart-cities-answer-population-fewer-resources (accessed May 20, 2021).

[34] X. G. Li and Q. M. Li, ‘THE APPLICATION OF DATA MINING TECHNOLOGY IN REAL ESTATE MARKET PREDICTION’, p. 6, 2006.

[35] D. Faggella, ‘Machine Learning in Real Estate - Trends and Applications Presented’, *Emerj*. https://emerj.com/ai-sector-overviews/machine-learning-in-real-estate-trends-and-applications/ (accessed May 21, 2021).

[36] D. Djenouri, R. Laidi, Y. Djenouri, and I. Balasingham, ‘Machine Learning for Smart Building Applications: Review and Taxonomy’, *ACM Comput. Surv.*, vol. 52, Feb. 2019, doi: 10.1145/3311950.

[37] ‘Smart building sensors for workplace analytics’, *PointGrab*. https://www.pointgrab.com/ (accessed May 21, 2021).

[38] ‘BuildingIQ – Building Energy and Operational Intelligence’. https://buildingiq.com/ (accessed May 21, 2021).

[39] O. Barzilay, ‘Property Management May Be The Next Frontier For AI’, *Forbes*. https://www.forbes.com/sites/omribarzilay/2017/03/14/property-management-may-be-the-next-frontier-for-ai/ (accessed May 21, 2021).

[40] ‘Commercial Real Estate Platform’, *VTS*. https://www.vts.com/ (accessed May 21, 2021).

[41] ‘Property Management Software | AppFolio’. https://www.appfolio.com/ (accessed May 21, 2021).

[42] ‘Zenplace | Making Rentals Easy for Landlords and Property Owners| Online Software Platform for Rental Properties’. https://www.zenplace.com/ (accessed May 21, 2021).

[43] ‘Zillow: Real Estate, Apartments, Mortgages & Home Values’, *Zillow*. https://www.zillow.com/ (accessed May 21, 2021).

[44] P. Martí, C. García-Mayor, and L. Serrano-Estrada, ‘Identifying opportunity places for urban regeneration through LBSNs’, *Cities*, vol. 90, pp. 191–206, Jul. 2019, doi: 10.1016/j.cities.2019.02.001.

Appendix

Example of appendix (extra figures, source code, etc.). Pay attention that reviewers are not forced to read the appendixes.