A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Santhosh Shanmugasundaram |
| *Student Number* | Sba24100 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | Rent Predictor: A Machine Learning Approach to Forecast Dublin Home Rent |
| *Assessment Due Date* | 2nd January 2025 |
| *Date of Submission* | 2nd January 2025 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Rent Predictor: A Machine Learning Approach to Forecast Dublin Home Rent

Contents

[1 Introduction 1](#_Toc186757493)

[2 Problem Definition 1](#_Toc186757494)

[3 Project Plan 1](#_Toc186757495)

[4 Business Understanding 2](#_Toc186757496)

[4.1 Objectives 2](#_Toc186757497)

[5 Data Understanding 3](#_Toc186757498)

[5.1 Collect Initial Data 3](#_Toc186757499)

[5.2 Describe Data 3](#_Toc186757500)

[5.3 Verify Data Quality 4](#_Toc186757501)

[5.4 Explore Data 5](#_Toc186757502)

[6 Data Preparation 6](#_Toc186757503)

[6.1 Select and Clean Data 6](#_Toc186757504)

[6.2 Construct and Format the Data 7](#_Toc186757505)

[7 Findings and Recommendations 8](#_Toc186757506)

[7.1 Modelling 10](#_Toc186757507)

[7.2 Train and Test 10](#_Toc186757508)

[7.3 Evaluation 11](#_Toc186757509)

[7.4 Plot Visualization 11](#_Toc186757510)

[8 Key Challenges 12](#_Toc186757511)

[9 Effort and Timeline 12](#_Toc186757512)

[9.1 Semester 1: 13](#_Toc186757513)

[9.2 Semester 2: 13](#_Toc186757514)

[10 Conclusion 13](#_Toc186757515)

[11 Future Development 13](#_Toc186757516)

[12 References 13](#_Toc186757517)

[13 GitHub Link 14](#_Toc186757518)

# Introduction

Rental homes are an integral part of the housing ecosystem, providing an alternative to homeownership. A properly functioning housing market is an essential ingredient of a properly functioning economy and society. (IPOA and IPAV)

From an economic, social and political perspective, housing is currently the greatest challenge faced by Irish policymakers. The housing challenge is manifested in a lack of supply of owner-occupier and rental properties; prohibitively high house prices and rents; and a serious problem of homelessness. (IPOA and IPAV)

# Problem Definition

The inability to provide an adequate supply of suitable and affordable housing supply for those who want to rent has very negative economic and social consequences. These consequences include:

* High and rising rents take spending power out of the economy and render it very difficult for aspiring house buyers to build up a sufficient deposit.
* High and rising house rents put upward pressure on wages, and this undermines national competitiveness.
* The availability of an abundant supply of high-quality housing to rent or purchase at affordable prices is a necessary condition for labour mobility within a country and between countries. For Ireland, inward migration is an essential part of the economic model, and housing can act as a major impediment to such labour flows.  
  (IPOA and IPAV)

# Project Plan

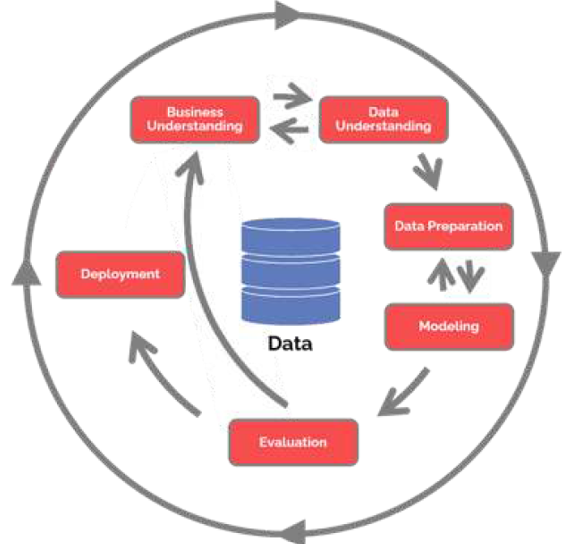
The scope of this project is to understand the **Project Management** strategy using key factors like increase in Dublin home rent price every quarter for three different types of homes namely apartment, terrace house and semi-detached house. The target of this study is to help the renters to decide on the future rent prices in Dublin so that expenses are planned.

Use the AI technology along with the Machine Learning models like ARIMA and Random Forest algorithm and predict the future rent at Dublin based on the **CRoss Industry Standard Process for Data Mining** (CRISP-DM).

This is a process model that serves as the base for a data science process that includes 6 phases as below.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

(Hotz)



# Business Understanding

The main objective of this phase is to list out the objectives and requirement of the project. This involves determining the business objectives, accessing the situation, determining the project goals and producing a project goal.

## Objectives

1. **Identify** the trend of home rent in Dublin using a Machine learning model with an appropriate dataset.

2. **Investigate** the important factors that affect the rental price.

3. **Develop** a model that can validate the existing dataset and plot different graphs depending on the number of bedrooms and type of house.

4. **Predict and forecast** future rental price based on the type of house and number of bedrooms.

# Data Understanding

In this step, we focus on identifying, collecting and analysing the dataset that can help this project. For this purpose, we investigate the Residential Tenancies Board which is an independent public body that regulates the private rented sector. It is responsible for multiple activities like tenancies registration, dispute resolution, investigation and sanction of landlords. It has also published many numbers of reports and data. (Citizensinformation.ie)

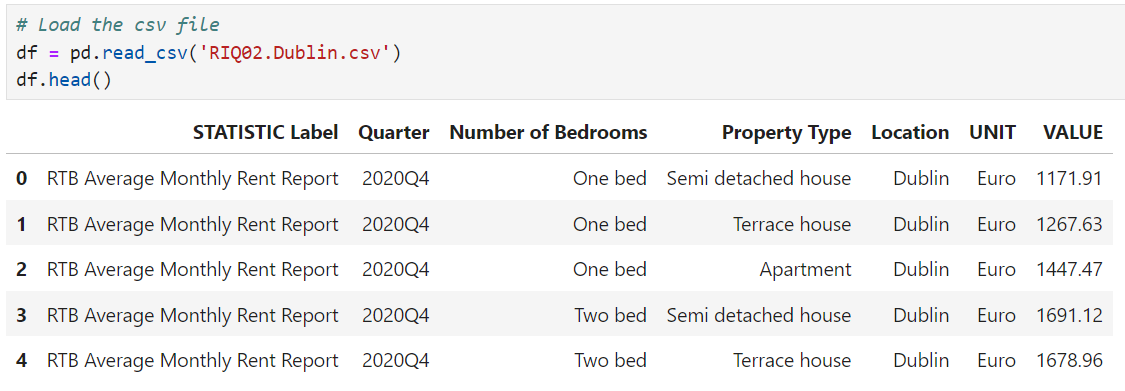
For the specific case study, I will use the dataset available in the below link.

<https://data.cso.ie/table/RIQ02>

## Collect Initial Data

The relevant data is downloaded in the ‘CSV’ file and uploaded to the python programming module using library named ‘PANDAS’.

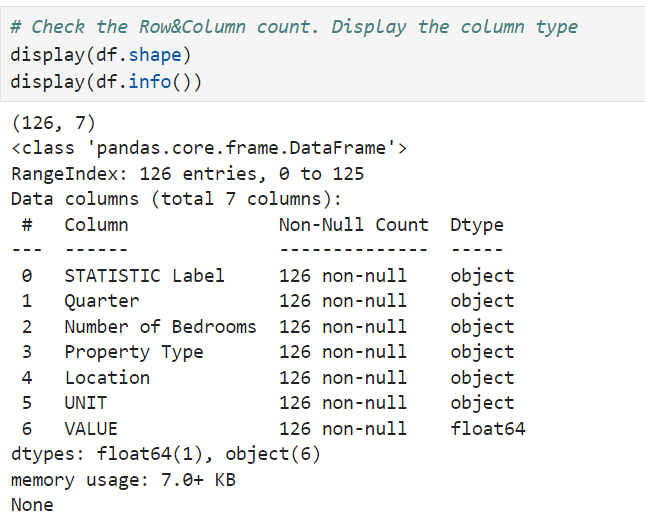
Once the data is loaded, the same can be displayed as-is using data frame function namely ‘HEAD’.



## Describe Data

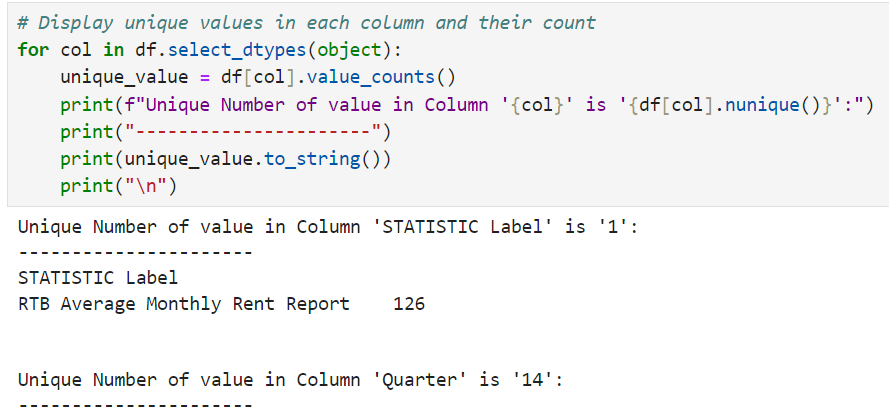
As part of the process, we examine the data and document its surface properties like data format, number of records etc.

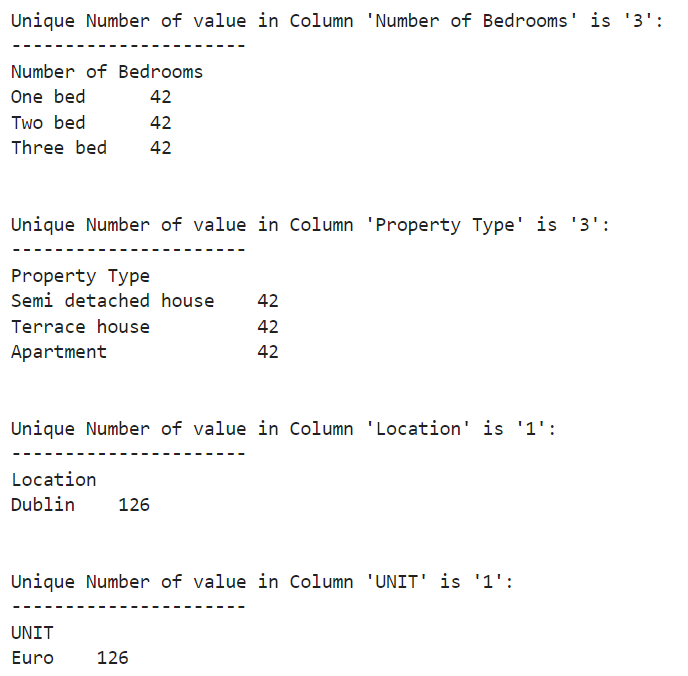
To check the number of rows and columns in the dataset, we use the data frame built in parameter ‘shape’ and to display the format of each data, we use the data frame function ‘info’.



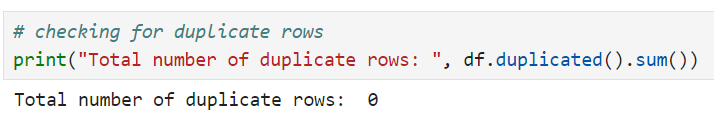
## Verify Data Quality

It is important to understand and inspect the data quality. To do that, we need to check the unique values in each of the column and check for any missing or duplicate records in the dataset.





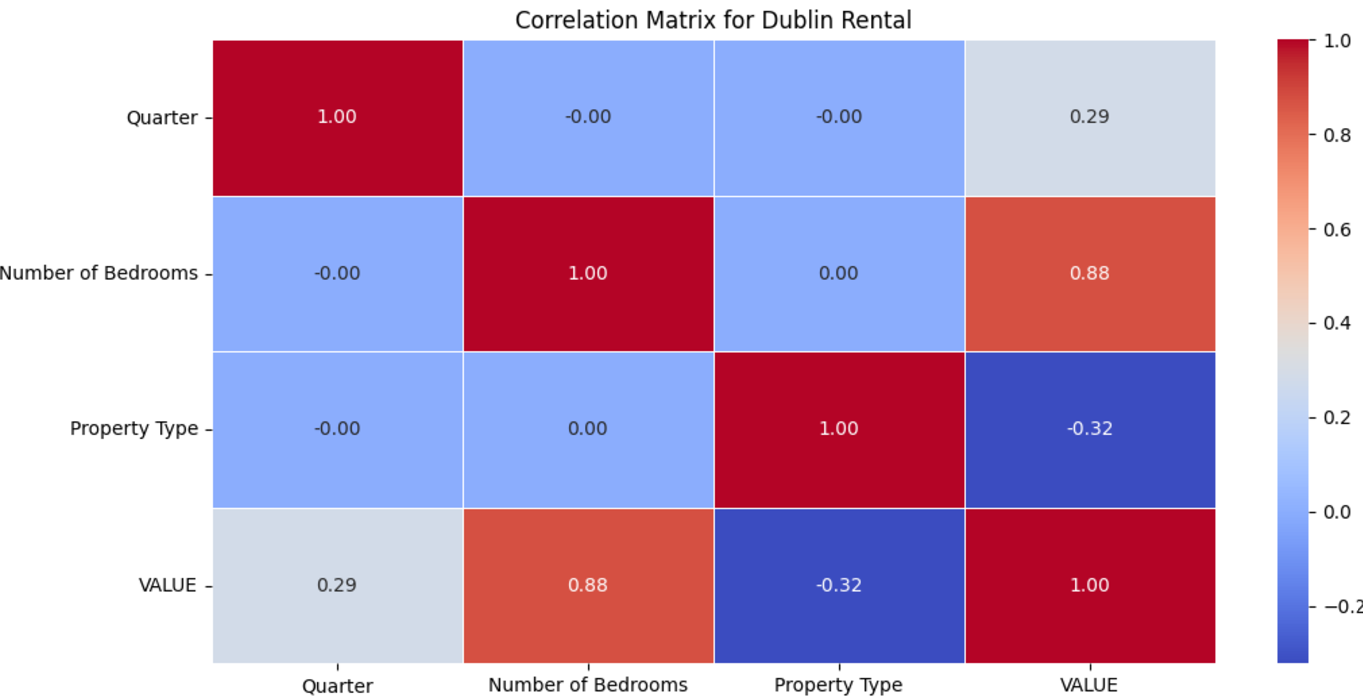
The above screen shot taken from jupyter notebook shows all the unique values corresponding to each column in the datasheet. This provides insight into what kind of data to expect in the outcome.



We must make sure that there are no duplicated entries in the dataset because this affects the model results. Thereby, we check for it and delete if any exists.

## Explore Data

As part of data understanding phase, it is critical to look deeper into data and draw a relationship between each column. To perform this, we derive the correlation factor associated with each column and find the most important data controls the result.



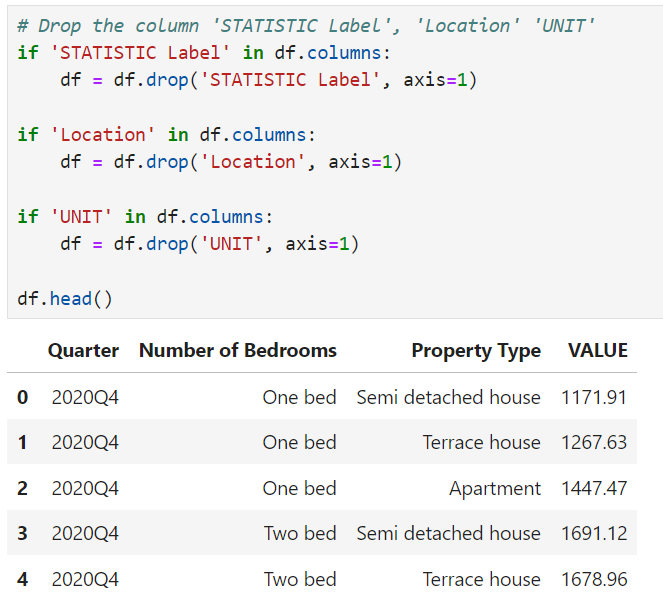
# Data Preparation

This phase which is often referred as ‘data mugging’, prepares the final dataset for modelling. A common rule of thumb is that 50% to 80% of project effort is spent in this phase. This phase has five important tasks namely,

1. Data selection
2. Clean data
3. Construct data.
4. Integrate data.
5. Format data.

## Select and Clean Data

A dataset may contain lot of information that may not help a machine learning algorithm in predicting the result. Therefore, it is important to find the columns that constitute no meaningful information and remove it. As a rule of thumb, any column that contains just 1 value for all records adds no value to the modelling. Therefore, such columns can be removed using data frame function ‘drop’.



For the dataset used in this project, I have removed 3 columns that has no real use for machine learning and the final columns that are relevant has been displayed and shown the screen shot above.

## Construct and Format the Data

It is very common for a dataset to contain information in strings and text. But for a machine learning model to work, it is very important to provide all the information in numeric format. As a common approach, we replace the unique text values into a simple meaningful number and same has been done as shown in the below screen shot.



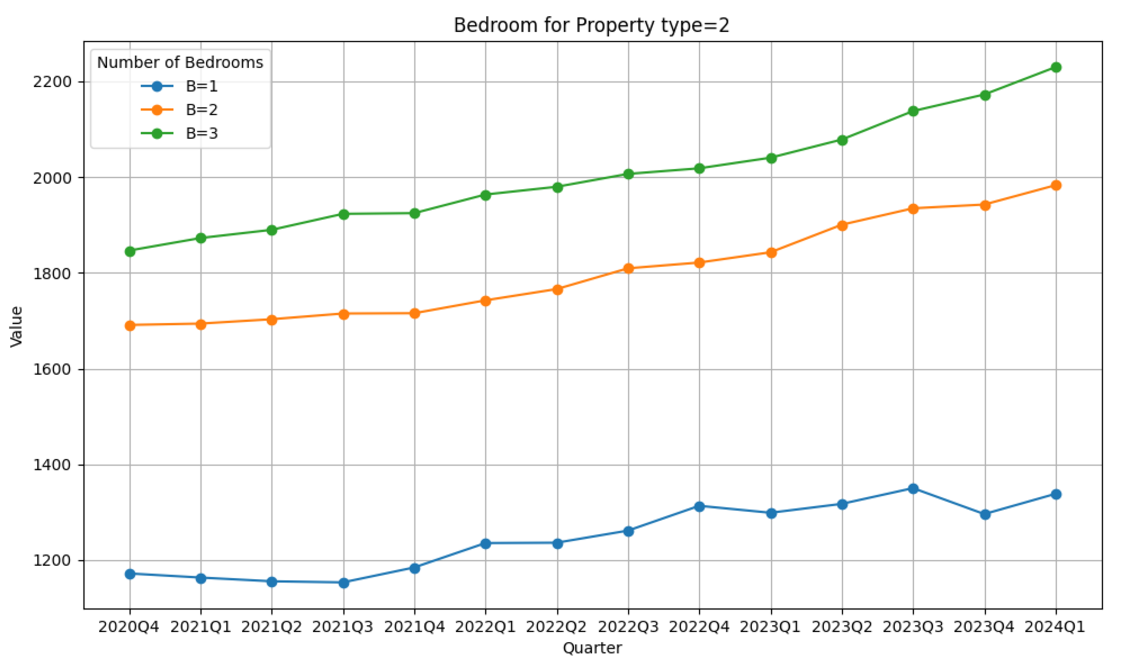
In this project, I have converted the number of bedrooms and property type into simple 0 to 3 number and same has been displayed above.

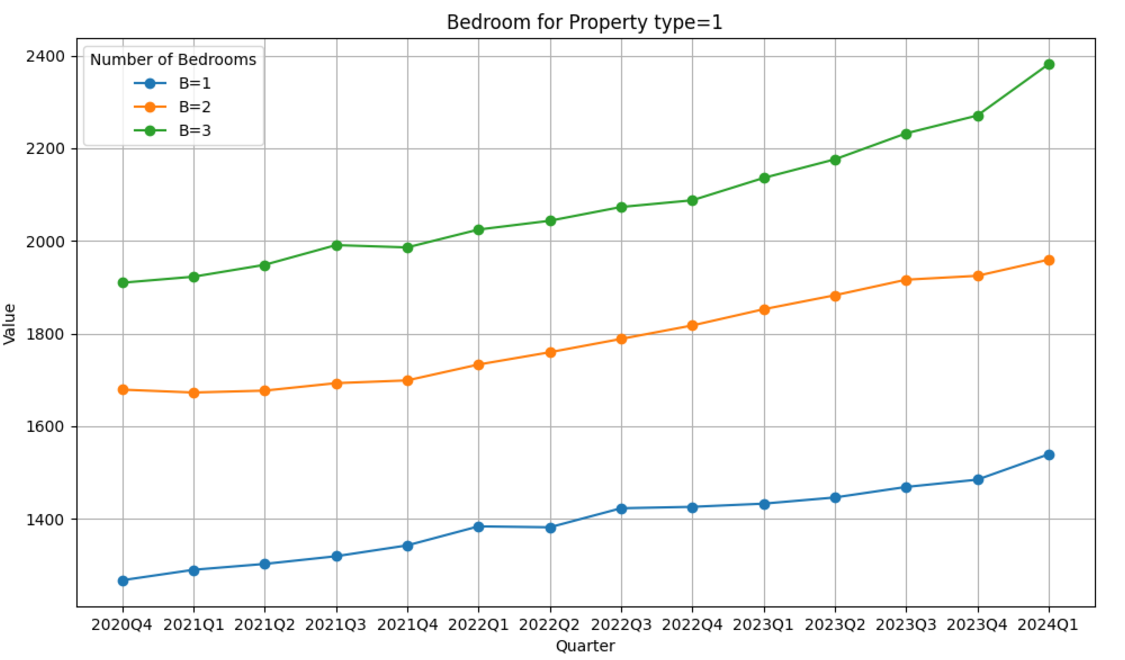
From now on, above naming conventions apply to the values in column ‘number of bedrooms’ and ‘property type’.

# Findings and Recommendations

Before we start on building a machine learning model, it is important to get a holistic view of the complete dataset and to do that we draw plots using library ‘matplotlib’.

For this project, I have created an individual plot based on the type of property and each plot will contain 3 graphs corresponding to bedroom numbers 1, 2 and 3. The X-axis corresponds to quarterly timeline and y-axis corresponds to the rental amount in Euro.







## Modelling

**ARIMA** (AutoRegressive Integrated Moving Average) is a very much famous and more powerful time series forecasting model which is used to investigate and predict time series data. It is very useful for data particularly dealing that exhibits trends and forecasting data. It combines the components of autoregression (AR), differencing (I), and moving average (MA) into a single model.

## Train and Test

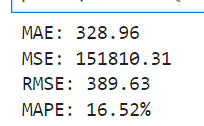
To train and test model, the actual data is split into 80% for training and 20% for testing. Since this project is about time series forecasting, it is important to split the data in continuous order so that data is not shuffled.

Once the data is split, using ARIMA model fit and forecast function, data is trained and forecasted.   
The forecasted data is then validated against the actual using below metrics,

1. Mean Absolute Error (MAE)  
   The average of the absolute differences between actual and predicted values.
2. Mean Squared Error (MSE)  
   The average of the squared differences between actual and predicted values. Larger errors have more impact.
3. Root Mean Squared Error (RMSE)  
   The square root of the MSE, providing error in the same units as the target variable.
4. Mean Absolute Percentage Error (MAPE)  
   The percentage error relative to the actual values.  
   (IBM)

## Evaluation

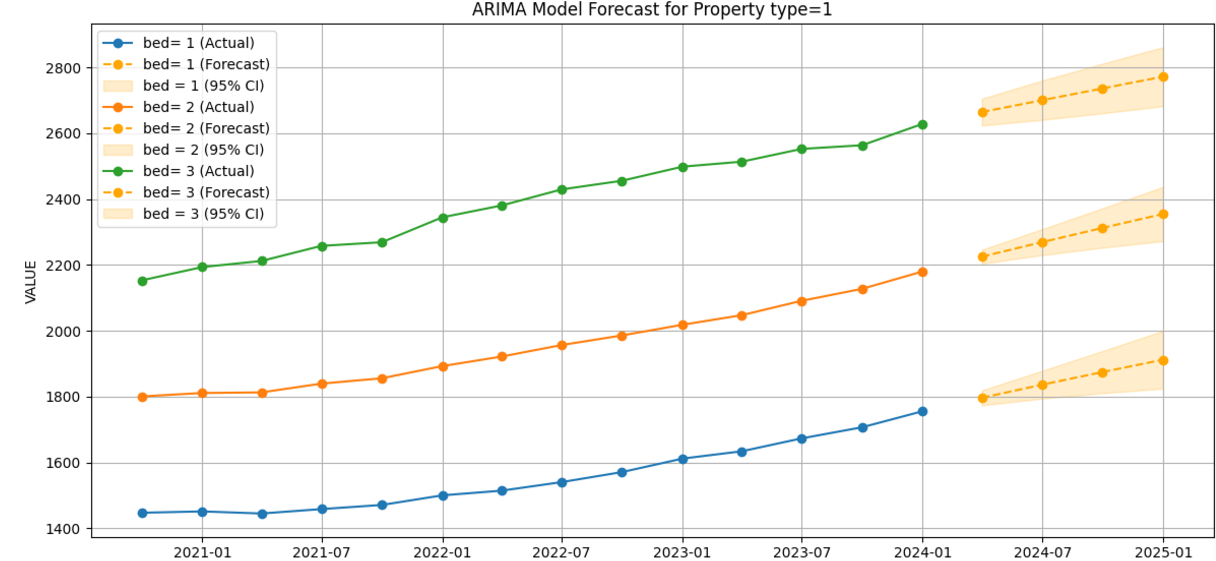
Following results has been obtained which tells that predicted value by the model is deviating from actual by 16.52%. Anything under 20% percentage is a good result.

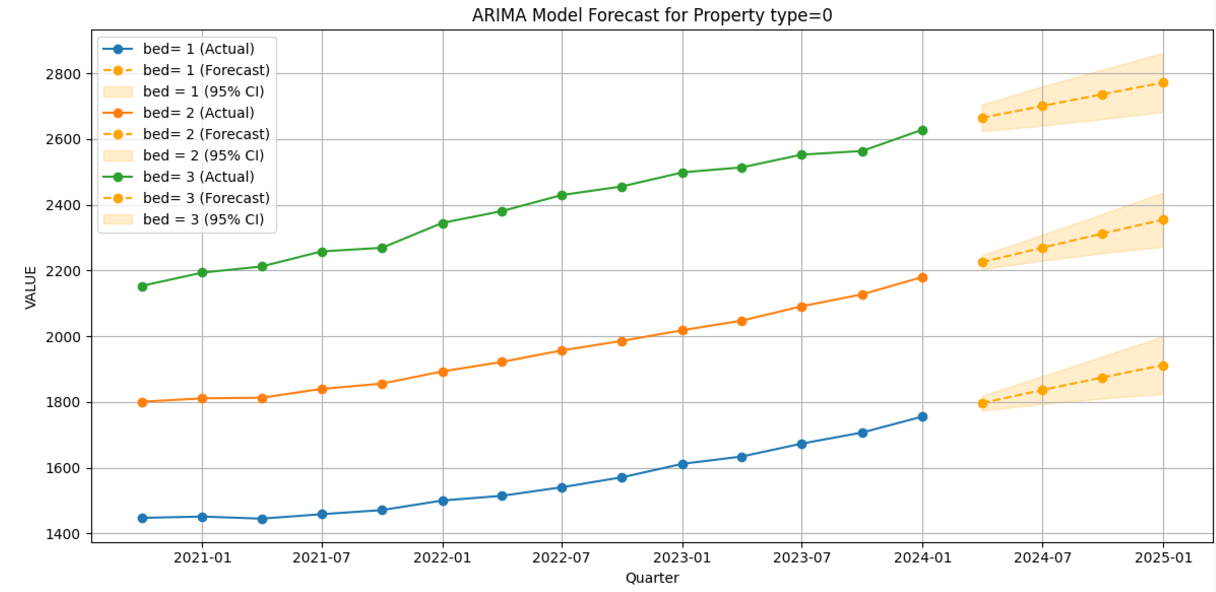


## Plot Visualization

Using the library ‘MATPLOTLIB’ following plots has been created.







# Key Challenges

Above mentioned dataset has some redundant and incomplete data which pose the treat of inefficiency during model training. Therefore, the key challenges are to segregate the useful and valid data from the large dataset, use the right filtering condition and at the same time not to miss on the important parameters.

# Effort and Timeline

Below provided efforts and timelines is an approximate estimate subject to change due to any unforeseen challenges.

## Semester 1:

|  |  |
| --- | --- |
| Activity | Duration |
| Identify and evaluate the dataset | 2 weeks |
| Perceive the possible challenges | 1 week |
| Model selection, training and testing | 2 weeks |
| Plots and graphs | 2 weeks |

## Semester 2:

|  |  |
| --- | --- |
| Activity | Duration |
| Data cleaning | 4 weeks |
| Complex modelling | 3 weeks |
| Graphical presentation | 1-2 week |
| Documentations | 2 weeks |

# Conclusion

According to various facts and references discussed above shows there is significant increase in the home rental fares in Dublin for the next 4 quarter. The upward trend in home rental shows high demand for housing and consumers must take prior care before considering relocating.

# Future Development

Now, prediction has been done only using ARIMA models but there is still scope for improvement by using ARIMA with other algorithms like Random Forest. At the same time, there is still more data that corresponds to specific areas in the Dublin city which shall be considered in the future development.

# References

(IPOA and IPAV)

IPOA, and IPAV. *THE IRISH PRIVATE RENTAL MARKET*. IPAV.ie, June 2022.

Available at: <https://www.ipav.ie/sites/default/files/ipav_ipoa_jim_power_updated_report_june_2022.pdf>

(Hotz)

Hotz, Nick. “What Is CRISP DM?” *Data Science Project Management*, 2024, www.datascience-pm.com/crisp-dm-2/.

Available at:

<https://www.datascience-pm.com/crisp-dm-2/>

(IBM)

IBM, IBM. “IBM Cognos Analytics 12.0.x.” *Ibm.com*, 18 Jan. 2024, www.ibm.com/docs/en/cognos-analytics/12.0.0?topic=forecasting-statistical-details.

Available at:

<https://www.ibm.com/docs/en/cognos-analytics/12.0.0?topic=forecasting-statistical-details>

# GitHub Link

<https://github.com/CCT-Dublin/ca1-capstone-project-proposal-santhosh-sba24100>