



كلية الحاسبات وتقنية المعلومات
College of Computers & Information Technology



Hadhramaut University
College of Computers and Information Technology
Computer Science Department

Real Estate Rental Advisor

A graduation project document submitted to the Dept. of Computer Science as partial fulfillment for the requirement for the Degree of B.Sc. in Computer Science

By

Abdulrahman Saeed Al-Dwani
Mohammed Abdullah Bagowabair
Mohammed Saleh Batook
Mohannad Khaled Alattass

Supervised by

Dr. Mohammed Abdullah Bamatraf

August, 2022

Abstract

Predictive models for deciding the estimated rental of Real estate in metropolitan cities is still remaining as more challenging and trickier task. The estimated rental of properties depends on a variety of interdependent factors. Key factors, which may affect the Real estate estimated rental, include area of the property, location of the property and its amenities. In this system, an attempt has been made to construct a predictive model for evaluating the estimated rental based on the factors that affect the Real estate estimated rental. Modelling study apply some supervised learning techniques such as Bayesian classifier or KNN algorithms. Such models are used to build a predictive model, and to pick the best performing model by performing a comparative analysis on the predictive errors obtained between these models. Here, the attempt is to construct a predictive model for evaluating the estimated rental based on factors that affects the Real estate estimated rental. this concept has built as real time application useful for real estate business and also buyer and sellers.

Dedication

We are so grateful for those who sacrificed their times for bringing us up ...

For those who gave us their most life to make us live and grow up ...

We are totally thankful for those who smiled a lot regardless the pain they have to make us so happy in our lives ...

To whom taught me the meaning of life ... To whom hold my hand on their paths
He is my teacher and my lover advises me if I make a mistake ... and holds my
hand if I stumble

... Dear Father ...

To the one who is pleased with my heart and blessed to the kindergarten of love
that sprouts flowers blossoms

... Dear Mother ...

Great thanks to those angels who were with us side by side all the times, we will
never ever forget their greatest favor...

Acknowledgment

We deeply thank Allah Al-Mighty for the blessing and success he granted us as well as we send our sincere thanks to the staff of our college

(The College of Computers and Information Technology)

Special Thanks and Grateful to

Dr. Mohammed Bamatraf

Dean of the College of Computers and Information Technology and Project
Supervisor

Special Thanks and Grateful to

Teacher. Khaled Bawazeer

Head of the Department of Computer Science

Thanks for everyone who provided the assistance and support to our project

Table of Contents

| Title | Page No. |
|-----------------------------------------------------------|-----------------|
| Abstract | i |
| Dedication | ii |
| Acknowledgment | iii |
| Table of Contents | iv |
| List of Figures | vii |
| List of Tables | ix |
| List of Abbreviations | x |
| Chapter 1: Introduction | 1 |
| 1.1 Introduction | 2 |
| 1.2 Problem Statement | 2 |
| 1.3 Objectives | 2 |
| 1.3.1 Main Objective | 2 |
| 1.3.2 Sub objectives | 3 |
| 1.4 Project Scope and Boundaries | 3 |
| 1.5 Methodology | 3 |
| 1.6 Tools used in the project | 4 |
| 1.6.1 Documentation Tools | 4 |
| 1.6.2 Software and Programming Languages | 4 |
| 1.6.3 Python Libraries & Frameworks | 4 |
| 1.7 Project Organization | 4 |
| Chapter 2: Background | 6 |
| 2.1 Introduction | 7 |
| 2.2 Background | 7 |
| 2.2.1 Real Estate | 7 |
| 2.2.2 Impact of Technology in Real Estate Industry | 7 |
| 2.2.3 Machine learning | 8 |
| 2.2.4 Data science | 9 |
| 2.2.5 Types of Real Estate Systems | 9 |

| Title | Page No. |
|--------------------------------------------------------------------|-----------------|
| 2.2.5.1 Real Estate Management Systems | 9 |
| 2.2.5.2 Real Estate Price Prediction Systems | 9 |
| 2.3 Related Work | 10 |
| 2.3.1 Bangalore House Price Prediction | 10 |
| 2.3.2 House price prediction in the UK | 10 |
| 2.3.3 Amlaki program | 10 |
| 2.3.4 Real estate manager program | 11 |
| 2.3.5 Falcon Pro program | 12 |
| 2.3.6 Fikra Program | 14 |
| 2.3.7 Online real estate | 14 |
| 2.3.8 REALas | 15 |
| 2.3.9 Zillow | 16 |
| Chapter 3: Data Collection & Preparation & Cleaning | 19 |
| 3.1 introduction | 20 |
| 3.2 Data | 20 |
| 3.3 Data Description | 21 |
| 3.4 Data preparation & cleaning | 30 |
| 3.4.1 Remove Duplicate | 30 |
| 3.4.2 Handel Missing Values | 31 |
| 3.4.3 Filter unwanted outliers | 33 |
| 3.5 Final prepared data | 34 |
| Chapter 4: Exploratory Data Analysis | 36 |
| 4.1 Introduction | 37 |
| 4.2 Descriptive Statistics | 38 |
| 4.3 Data Distribution and Visualization | 41 |
| 4.4 Correlation | 60 |
| 4.4.1 Correlation Matrix | 60 |
| 4.4.2 Analysis of variance (ANOVA) | 62 |
| Chapter 5: Experimental and Discussion Modeling | 71 |
| 5.1 Introduction | 72 |

| Title | Page No. |
|----------------------------------------------|-----------------|
| 5.1.1 Confusion Matrix | 72 |
| 5.1.2 Classification Measure | 74 |
| 5.2 Experiment and result | 76 |
| 5.2.1 Experiment Models | 76 |
| 5.2.2 Result | 81 |
| Chapter 6: Model Implementation | 85 |
| 6.1 introduction | 86 |
| 6.2 MLP implementation | 88 |
| 6.3 MLP visualization | 92 |
| Chapter 7: Conclusion and Future work | 93 |
| 7.1 Introduction | 94 |
| 7.2 Conclusion | 94 |
| 7.3 Future work | 94 |
| References | 95 |
| الخلاصة بلعربي | 97 |

List of Figures

| Figure | Page No. |
|----------------------------------------------------------------------------------------------------------------------|----------|
| Figure (1.1) Methodology | 4 |
| Figure (2.1) Amlaki program | 11 |
| Figure (2.2) Real estate manager program..... | 12 |
| Figure (2.3) Falcon Pro program..... | 13 |
| Figure (2.4) Fikra Program..... | 14 |
| Figure (2.5) Online real estate..... | 15 |
| Figure (2.6) REALas | 16 |
| Figure (2.7) Zillow..... | 17 |
| Figure (3.1) Number and locations of data collected..... | 21 |
| Figure (3.2) Data collected from marine side and mountain side | 22 |
| Figure (3.3) Residential/Commercial Real Estate..... | 24 |
| Figure (3.4) number of the deluxe and standard property..... | 25 |
| Figure (3.5) Describe the types of real estate contracts collected..... | 25 |
| Figure (3.6) Represent the number of concrete and non-concrete property. | 26 |
| Figure (3.7) Description of the rental period for old tenants for each area. | 28 |
| Figure (3.8) Access road type stats for each region | 29 |
| Figure (3.9) Property condition for each area | 29 |
| Figure (3.10) Delete duplicates by excel..... | 30 |
| Figure (3.11) Example of data before filling in the missing records | 31 |
| Figure (3.12) Example of data after filling in the missing records | 32 |
| Figure (3.13) Example of data before deleting lost records..... | 32 |
| Figure (3.14) Example of data after deleting lost records..... | 33 |
| Figure (3.15) Clarification of data before deleting outliers | 34 |
| Figure (3.16) Clarification of data after deleting outliers | 34 |
| Figure (3.17) Aparentemnts after clening | 35 |
| Figure (4.1) Snapshot of the dataset..... | 38 |
| Figure (4.2) Data Information. | 39 |
| Figure (4.3) Count, mean, STD, min, max, 25%, 50% and, 75% information for Price and Distance | 40 |
| Figure (4.4) Most frequent value in each column of the data. | 41 |
| Figure (4.5) Data distribution is clarified in terms of the property's status, whether it is standard or deluxe..... | 42 |
| Figure (4.6) Distribution of data on categories from 0 to 4 in N-balconies column | 43 |
| Figure (4.7) Distribution of data on categories from 0 to 3 in the N-bathrooms column | 43 |
| Figure (4.8) Distribution of data on categories from 0 to 4 in the kitchens column | 44 |

| | |
|--------------------------------------------------------------------------------------------------------------------------------------------------|--------|
| Figure (4.9) Distribution of data on categories from 0 to 4 in the halls column | 44 |
| Figure (4.10) Distributed of data to five categories, east, west, south, north, and open. | 45 |
| Figure (4.11) Type of access road..... | 46 |
| Figure (4.12) Space categories. | 46 |
| Figure (4.13) Distribution of data on categories from 1 to 6 in the number of rooms column | 47 |
| Figure (4.14) Number of properties whose construction type is concrete and the number of properties whose construction type is non-concrete..... | 47 |
| Figure (4.15) Families / Singles attribute..... | 48 |
| Figure (4.16) Mountainous and marine Categories | 49 |
| Figure (4.17) Neighborhood Caegories..... | 49 |
| Figure (4.18) Population-density catgrorice | 50 |
| Figure (4.19) Property-type catgrorice..... | 50 |
| Figure (4.20) Property-Age categories..... | 51 |
| Figure (4.21) Property Condition categories..... | 51 |
| Figure (4.22) Rental-period categories | 52 |
| Figure (4.23) Residential/Commercial categories | 52 |
| Figure (4.24) Services categories. | 53 |
| Figure (4.25) Street-type categories. | 53 |
| Figure (4.26) Water-meter categories. | 54 |
| Figure (4.27) Adjacent-Sides categories. | 54 |
| Figure (4.28) Contract categories..... | 55 |
| Figure (4.29) N-enterances categories. | 55 |
| Figure (4.31) N-floor categories..... | 56 |
| Figure (4.32) Price distribution histogram. | 57 |
| Figure (4.33) Price Boxplot..... | 58 |
| Figure (4.34) Distance distribution histogram. | 59 |
| Figure (4.35) Distance Boxplot..... | 60 |
| Figure (4.36) Price and Distance Correlation | 61 |
| Figure (4.37) Correlation matrix heatmap | 61 |
| Figure (5.1) 2x2 Confusion Matrix | 73 |
| Figure (5.2) Data incorrectly/correctly classified in MLP alogorithem..... | 83 |
| Figure (5.3) The average recall rate and FP rate for each algorithm | 83 |
| Figure (5.4) F-Measuer and Precision for each algorithm | 84 |
| Figure (6. 1) layers perceptron | 85 |
| Figure (6. 2) Input Widegets. | 85 |
| Figure (6.3) MLP visualization | 85 |

List of tables

| | |
|--------------------------------------------------------------------------------|----|
| Table 3.1 Number and locations of data collected..... | 21 |
| Table 3.2 Residential/Commercial Real Estate..... | 24 |
| Table 3.3 Property condition for each area..... | 29 |
| Table 5.1 summarizes the result of the comparison and evaluation process | 81 |

List of Abbreviations

| <u>Abbreviation</u> | <u>Meaning</u> |
|----------------------------|---------------------------------------------------------|
| AI | Artificial Intelligence |
| ML | Machine Learning |
| DS | Data Science |
| REMS | Real Estate Management System |
| KNN | K-Nearest Neighbor |
| EDA | Exploratory Data Analysis |
| STD | Standard Deviation |
| ANOVA | Analysis of Variance |
| IQR | Inter Quartile Range |
| H0 | Null Hypothesis |
| TP | True Positive |
| TN | True Negative |
| FP | False Positive |
| FN | False Negative |
| LWL | Locally weighted learning |
| Rep | Reduced Error Pruning Tree |
| JRIP or RIPPER | Repeated Incremental Pruning to Produce Error Reduction |

Chapter One

Introduction

1.1 Introduction

The area of Real Estate is important to individuals and stakeholders, with increasing of the growth and rapid economic development, estimated rental change dramatically from time to time.

The estimated rental of Real Estate are affected by many different factors, which includes geographic location, economic and physical conditions of the individual and the region, with the present of many systems that used to manage real estate, most of these systems serve stakeholders and landlords in determining the available property, geographic location, pricing, and using in financial operations such as inventory, annual earnings determination, and other financial operations, but all of these systems has lack of predictions and are standard archiving systems.

Real Estate Rental Advisor provides highly accurate rental price predictions for shops and apartments using machine learning and data science to improve and predict prices with high accuracy.

The real estate rental advisor system can learn from a large set of data to teach itself to optimize parameters and make data-driven predictions. In order to obtain accurate results, a large amount of data was collected in many different areas, targeting the Fouth area in particular, and work was also done to collect various data, including apartments and rented and for rent shops, to obtain a survey and accurate results in forecasting and testing process.

1.2 Problem Statement

- Relying on insufficient factors in determining real estate estimated rental.
- The need of investors and stakeholders to know the trend of real estate estimated rental in a particular location.

1.3 Objectives

The project will achieve the following objectives:

1.3.1 Main Objective

Develop a Real Estate Rental Advisor System employing machine learning.

1.3.2 Sub Objectives

The main objective will be done by the following sub objectives:

- Data collecting, cleaning and analyzing.
- Model the data with different algorithms.
- Evaluation the models.
- Implement the prediction model.

1.4 Project Scope and Boundaries

This System some information such as the location of a property to give a price prediction about the property, and it also provides some analysis about the real estate situation in the chosen area, which is easy to use and can be used by anyone (brokers or office owners) who wants to see the value of a particular property.

So far, the data collected inside Fouh from Ibn-Sina to Inshaat, and this will be the scope and limitations of the software.

1.5 Methodology

The Real Estate Rental Advisor system is used to make real estate estimated rental predictions such as apartments and stores. The process of creating the system goes through several steps, starting with defining the data requirements that were identified through interviews and field visits to real estate owners and real estate realtors according to what the system needs such as property type, price, etc., then the data collection process took place in the specified area and after collecting a large amount of data, it is cleaned up and excluded the unimportant and missing data, and then this data will be carefully analyzed and checked for possible logical groupings and hidden relationships using methods, graphics and statistics, in the process of developing the model, plots of the data, process knowledge and assumptions about the process are used to determine the form of the model to be fit to the data, then the

system will be implemented and created using the appropriate programming



Figure 1.1:Methodology

languages and tools.

1.6 Tools used in the project

In this project the following tools will be used.

1.6.1 Documentation Tools

- Microsoft Office 2019.

1.6.2 Software and Programming Languages

- Anaconda Individual Edition 2020.11.
- Tableau 2019 Enterprise Edition.
- Programming Languages (Python and JavaScript).

1.6.3 Python Libraries & Frameworks

- NumPy, Pandas, Seaborn, Matplotlib, SciPy, Stats, Sklearn, Iwidgets.

1.7 Project Organization

The project document is organized as the following chapters:

- Chapter 1 introduction: It contains an introduction about the project, the problems need to solve, the objectives need to achieve, the scope of the project, as well as the tools used to do the project.
- Chapter 2 Background & Related Work: It contains the basic concepts addressed by the project with a mention of previous projects related to the project and similar to it.
- Chapter 3 Data Collection & Preparing & Cleaning: It contains a description of the data, the attribute and the results that obtained after collecting the data, and the process of cleaning the data with clarification of the results after cleaning.
- Chapter 4 Exploratory Data Analysis: In this chapter, the data are analyzed, graphically illustrated, processed and extract actionable and relevant information that helps make informed decisions and predictions.
- Chapter 5 Experimental and Discussion Modeling: In this chapter, some algorithms are evaluated and compared to get the best.
- Chapter 6 Model Building & Implementation: It contains a description of the algorithm used to build the model and the code of that algorithm.
- Chapter 7 conclusion and future work: This chapter presents the results and conclusions and the future work.

Chapter Two

Background

2.1 Introduction

In this chapter, the basic concept in the field of real estate and its development and its relationship to technology will be explained, and the types of archival and predictive systems and how they work, with an explanation of procedures and concepts based on them, such as machine learning and data science, with real examples, explaining their work, usefulness and importance in the field of real estate.

2.2 Background

2.2.1 Real Estate

Real estate is the land along with any permanent improvements attached to the land, whether natural or man-made—including water, trees, minerals, buildings, homes, fences, and bridges. Real estate is a form of real property. It differs from personal property, which are things not permanently attached to the land, such as vehicles, boats, jewelry, furniture, and farm equipment. Despite the magnitude and complexity of the real estate market, many people tend to think the industry consists merely of brokers and salespeople. However, millions of people in fact earn a living through real estate, not only in sales but also in appraisals, property management, financing, construction, development, counseling, education, and several other fields [1].

2.2.2 Impact of Technology in Real Estate Industry

Real estate has always been a technologically oriented industry, but there have been some considerable advancements in the past few years. The development of real estate technology has had an impact on how the industry runs and has changed the face of the industry itself. These new technologies that are impacting the real estate industry will continue to impact the industry in 2022 and beyond. Whether a real estate agent looking to sell more homes, a landlord looking for better tenants, or a first-time homebuyer, these advancements are going to impact the life significantly. Undoubtedly, the growth of technology has resulted in substantial improvements in

the operation of online buying and selling platforms in recent years. Real estate operations and accessibility have become significantly simpler and more user-friendly in recent years. Making contact with prospective tenants or homeowners has become a breeze for landlords, as well as for tenants to reach out to landlords. Today, buying, selling, or even renting a property has become a cakewalk for everyone, unlike in the old days. Real estate agents can now easily cater to the demands. Whether a real estate agent is looking to sell a house or a landlord looking for an ideal renter, or a homeowner looking to resell his/her house, everything can be done with just a few clicks – all thanks to the technology in real estate! Whether it's for constructing, construction services, home services, acquiring, selling, or renting, the adoption of technology has drastically altered the interaction process for all parties involved in the transaction [2].

2.2.3 Machine learning

Machine learning (ML) is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Machine learning can be the key to unlocking the value of corporate and customer data, and enacting decisions that keep a company ahead of the competition [3].

ML is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow [4].

ML has proven valuable because it can solve problems at a speed and scale that cannot be duplicated by the human mind alone. With massive amounts of computational ability behind a single task or multiple specific tasks, machines can be trained to identify patterns in and relationships between input data and automate routine processes [5].

2.2.4 Data science

Data science is a multidisciplinary approach to extracting actionable insights from the large and ever-increasing volumes of data collected and created by today's organizations. Data science encompasses preparing data for analysis and processing, performing advanced data analysis, and presenting the results to reveal patterns and enable stakeholders to draw informed conclusions [6].

Data science is important for businesses because it has been unveiling amazing solutions and intelligent decisions across many industry verticals. The epic way of using intelligent machines to churn huge amounts of data to understand and explore behavior and patterns is simply mind-boggling. This is why data science has been getting all the spotlight [7].

2.2.5 Types of Real Estate Systems

There are two types of Real Estate Systems:

2.2.5.1 Real Estate Management Systems

Real Estate Management System (REMS) is a real estate software application that manages the overall operational activities and processes, starting from the management of the property, to the management of real estate agencies, agents, clients and financial transactions. It provides comprehensive reports for managing the Real Estate agency performance and efficiency [8].

2.2.5.2 Real Estate Price Prediction Systems

Modelling uses Machine Learning Algorithms, where machine learns from the data and uses them to predict a new data. The most frequently used model for predictive analysis is Bayesian classifier. The proposed model for accurately predicting future outcomes has applications in economics, business, banking sector, healthcare industry, e-commerce, entertainment, sports etc. One such method used to predict house prices are based on multiple factors. Predictive models for deciding the sale or rent price of properties in metropolitan cities is still remaining as more challenging and trickier task. The sale price of properties in cities depends on a

variety of interdependent factors. Key factors which may affect the house price include area of the property, location of the property and its amenities. Modelling study apply some supervised learning techniques such as Bayesian classifier or K-Nearest Neighbor (KNN) algorithms. Such models are used to build a predictive model, and to pick the best performing model by performing a comparative analysis on the predictive errors obtained between these models [9].

2.3 Related Work

2.3.1 Bangalore House Price Prediction

It is one of the research projects conducted in India in the city of Bangalore, which aims to implement a house price prediction model of Bangalore, India. It's a Machine Learning Model which integrates Data Science and Web Development. Housing prices fluctuate on a daily basis and are sometimes exaggerated rather than based on worth. The major focus of this project is on predicting home prices using genuine factors, to view the research, visit the website specified in the reference [10].

2.3.2 House price prediction in the UK

The project aims to answer the question of how some variables affect the change of property's prices over a long period of time. The dataset is an official record of all transactions recorded from 1995 to 2017 in the UK by the HM Land Registry, responsible for administrating the real estate, to view the research, visit the website specified in the reference [11].

2.3.3 Amlaki program

The program performs many operations related to real estate, managing it financially, collecting the rents owed to it and dealing with expenses on it, as well as acting as a property management company on behalf of the owner in all operations according to the contract, including the tasks required of following up on rents, cash sales, installments, follow-up of maintenance and services on real estate or accordingly for the agreed upon through the registered contract with the real estate owner on how to manage his property or to be an investment contract[12,13].

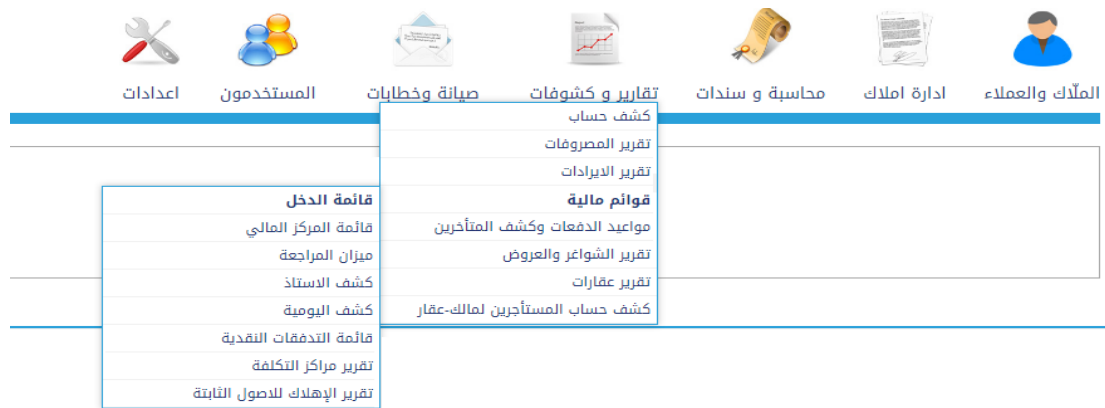


Figure 2.1:Amlaki program

Advantages:

- Financial and administrative property management.
- Easy to use.
- Provides an unlimited number of users.
- An integrated accounting and financial system and electronic reports.
- An integrated electronic messages and notifications system.
- Allows owners to follow up on their properties through the Internet.

Disadvantages:

- Unintelligent and it does not provide prediction for real estate prices.
- It does not support more than one language.

2.3.4 Real estate manager program

The real estate manager program is an integrated program for managing real estate, rents, installments and plans. It is also possible to print sales and purchase contracts and lease contracts through the program, and it deals with the Hijri or Gregorian dates [14].

بحث في الإيجارات

أولاً : قم بتحديد نوعية البحث من الخيارات أدناه.

☒ البحث باسم المستأجر
☐ البحث باسم المؤجر
☐ البحث بالموقع
☐ البحث بعدد دورات المياه
☐ البحث بالإيجار السنوي
☐ البحث برقم الهوية
☐ البحث بتاريخ بداية الإيجار
☐ البحث بتاريخ نهاية الإيجار

☐ البحث بهاتف المستأجر
☐ البحث بالنوع
☐ البحث برقم البناية
☐ البحث برقم العقد
☐ البحث بالجنسية
☐ البحث برقم الشقة

2007 / 5 / 15
 2007 / 5 / 15

ثانياً : أدخل كلمة البحث أدناه أو اخترها من القائمة ثم اضغط على زر البحث

أحمد حافظ عبد الرحمن

ثالثاً : لعرض أو تعديل أي بيانات في نتائج البحث قم بالضغط على الاسم أدناه ثم اضغط على زر (عرض / تعديل)

| م | اسم المستأجر | هاتف المستأجر | هاتف المؤجر | بداية الإيجار |
|---|----------------------|---------------|-------------|---------------|
| 1 | أحمد حافظ عبد الرحمن | - | - | 5/2/1428 |

خروج
 طباعة نتائج البحث
 عرض / تعديل

Figure 2.2: Real estate manager program

Advantages:

- The program works on real estate management, so it is possible to enter real estate data and determine its quality.
- Contains a comprehensive search engine and also can print property data.
- It is characterized by the management of rents. It is possible to enter rental data, specify the date and end of the lease, and specify the dates and amounts of installments, if available.

Disadvantages:

- Difficult to use.
- Unintelligent and it does not provide prediction for real estate prices.
- It does not support more than one language

2.3.5 Falcon Pro program

An integrated program to manage all real estate activities (sales - purchases - rents - clients - commissions - integrated accounts system - profit and loss account for each project - check management - police reports and lawsuits management - email and text message alerts and notifications).

The real estate program Falcon Pro is one of the best real estate programs available in Arabic and English [15].

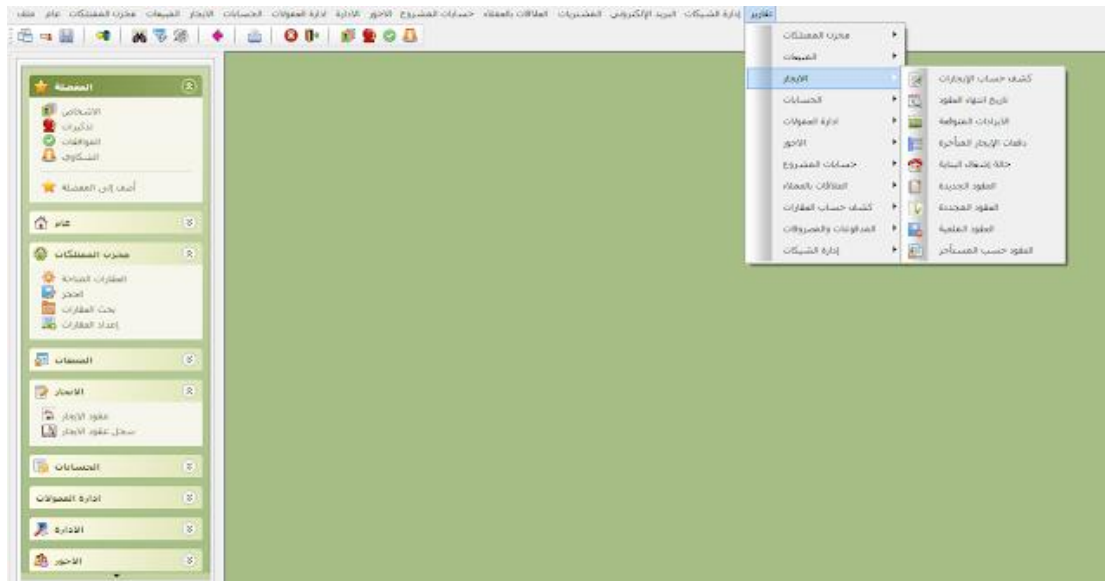


Figure 2.3:Falcon Pro program

Advantages:

- Configure the program according to the business needs and constantly update the program for free.
- The program was developed by experts in various real estate fields.
- Reports for each project / property, ease of calculating profits and losses, and managing dealing with clients through the program.
- The databases are located on the client's machines, which gives the highest degree of security.

Disadvantages:

- Unintelligent and it does not provide prediction for real estate prices.
- Difficult to use.

2.3.6 Fikra Program

Fikra Real Estate Program is the best real estate accounting program for real estate companies and offices in order to organize sales and rents, in addition to a database management system and real estate marketing system, in addition to a professional accounting system [16].

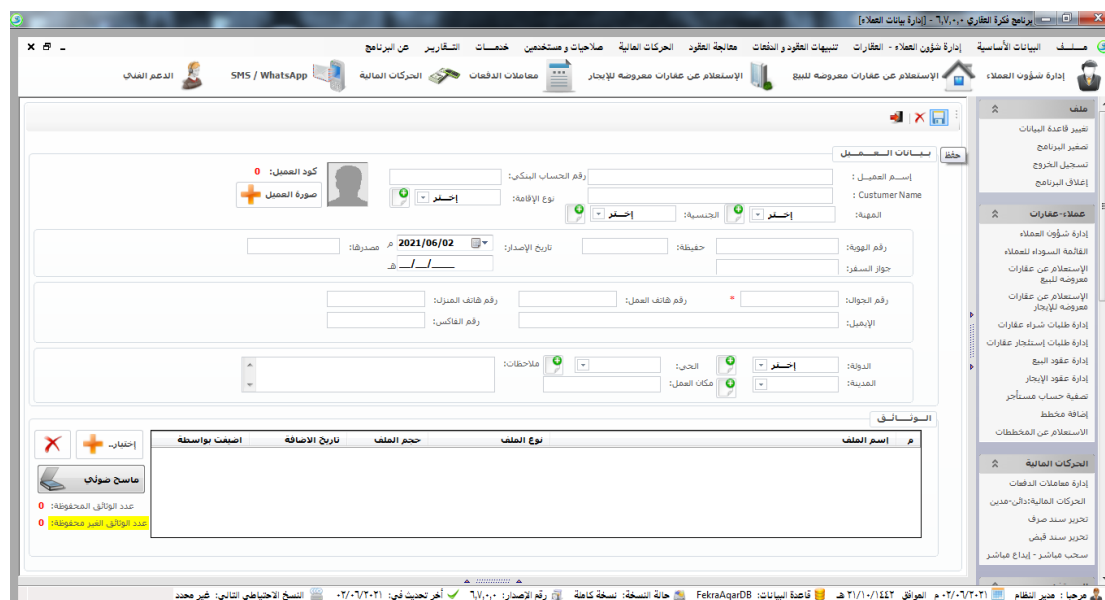


Figure 2.4:Fikra Program

Advantages:

- The property management program "Fikra" serves all real estate fields such as selling, renting, reports and securities, in addition to the real estate database management system and the distinguished real estate marketing system.

Disadvantages:

- Unintelligent and it does not provide prediction for real estate prices.
- High system cost.

2.3.7 Online real estate

A Yemeni site for buying and selling real estate in Yemen. The program facilitates the search for real estate and contact owners to inquire about the details of the property, view it on the ground, and complete the buying and selling process [17].

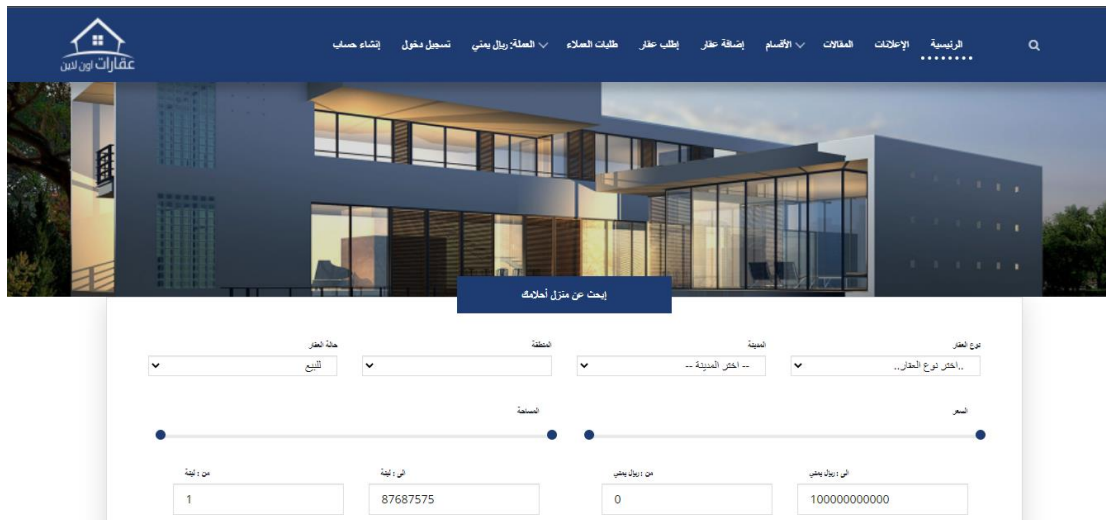


Figure 2.5:Online real estate

Advantages:

- Contains hundreds of properties.
- Adding real estate daily.
- Show the price in the three currencies (Yemeni rial, Saudi rial, US dollar).
- View property photos and details.
- Search with all property information (price, age, governorate, region, ... etc.).
- Adding a feature to make tabs for owners of real estate offices
- Adding the advertisement number to the advertisement images for easy reference.
- Enable access to articles.

Disadvantages:

- It does not provide predictions for real estate prices.
- It does not support more than one language.
- The area is not clearly defined.

2.3.8 REALas

It's a property price prediction tool that buyers can use to get a clearer idea of what a property is really worth. REALas processes local, historic and recent data through an algorithm to create a prediction of a property's sale price [18,19].

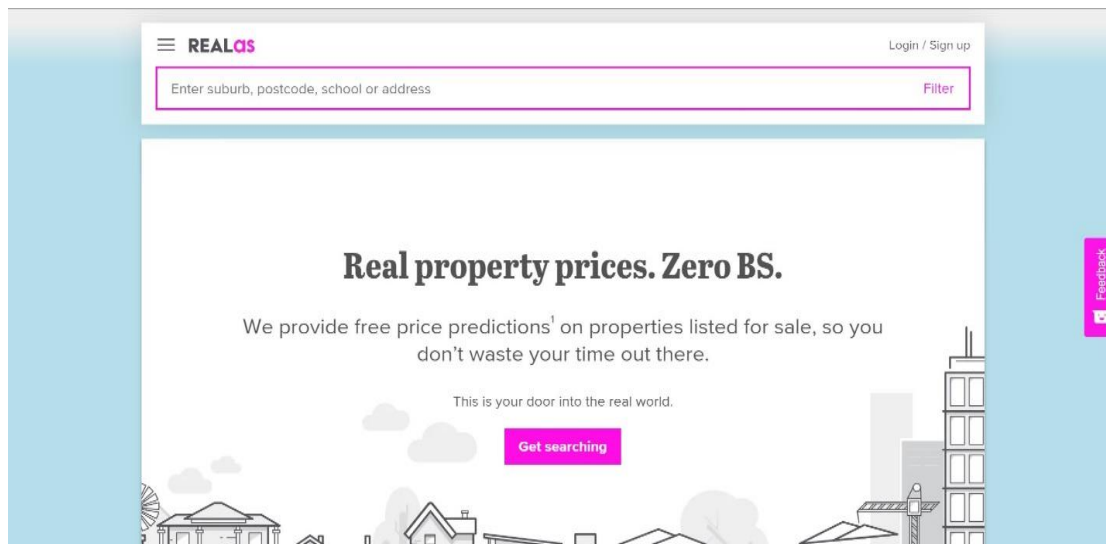


Figure 2.6:REALas

Advantages

- Don't waste time looking for unafforded properties.
- Provides predictions are accurate within 5% on 7 out of 10 properties and within 10% on 8.7 out of 10.
- Easy to get the Real Estate stats like how long the property has been on the market.
- Learn about a local market faster by tracking the properties, love, like or share them.

Disadvantages:

- REALas can only predict prices for residential properties, and it cannot predict prices of incomplete construction
- REALAS works exclusively in Melbourne and Australia.

2.3.9 Zillow

Zillow is a popular online real estate and rental marketplace dedicated to provide customers with data to make the best possible housing decision [20].

The Zillow Home Value prediction is based on a statistical model using a variety of economic data. The model takes into account economic and housing data that might have an impact on future home values [21].

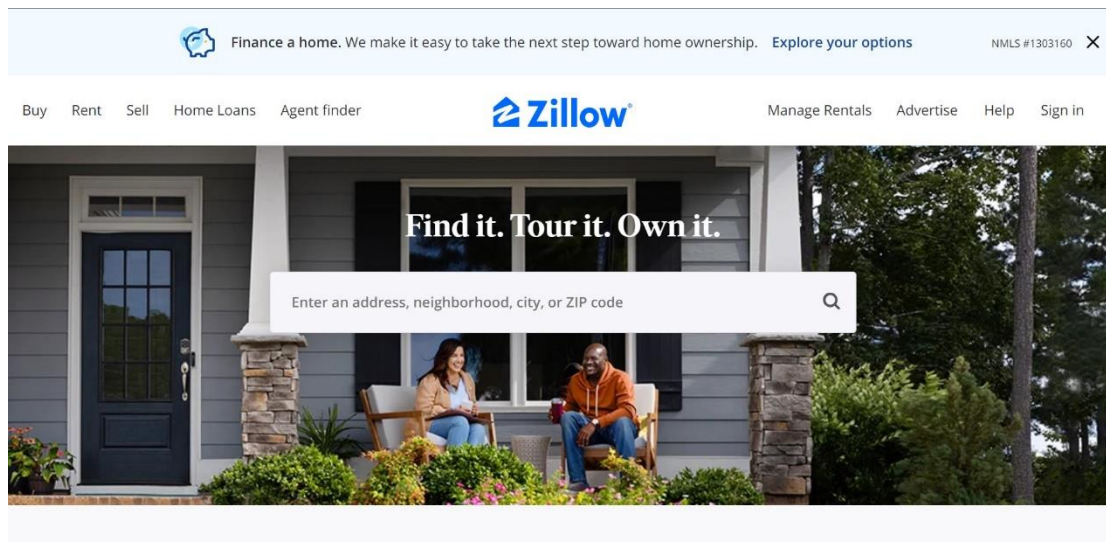


Figure 2.7:Zillow

Advantages

- Zillow offers what it calls Zestimates, which are estimates of home values based on publicly available information. Graphically, each site presents listings in a different way, which provides the user with a different experience.
- Zillow provides users with a highly graphic experience when searching for properties.
- Zillow is free to use for both owners, listing agents, and landlords.
- Add three-dimensional tours to get 360° photos of houses customers are interested in buying or renting.
- Zillow has added a tool for potential renters to submit a credit check and eviction history report to landlords.

Disadvantages:

- Its work is limited on the United States only.

It turned out that the previously mentioned systems were archival systems and there is no help for the customer in predictive matters, which may help the customer in making a decision such as determining the rental price, and that these systems are not considered unintelligent, and some other systems have been mentioned, which are smart and predictive, but it is limited to certain areas and cannot be predicted in other areas.

Chapter Three

Data Collection & preparation & cleaning

3.1 Introduction

In this chapter, the basic concept of data, the method of collecting and processing it, the process of cleaning it, and clarification of the results before and after the cleaning process were clarified.

3.2 Data

In general, data is any set of characters that is gathered and translated for some purpose, usually analysis. If data is not put into context, it doesn't do anything to a human or computer [22].

Data should reflect a fact or a group of facts in the real world, which is its foremost important characteristic. Data can be collected in different ways and stored in various formats, and the optimal way depends on usefulness and efficiency. There could be many attributes of a fact. It takes much more resources and time to collect all the details, while a subset may prove not enough in the end. The efficiency also depends on the data format and structure. Putting data into a numerical form in a quantity's manner takes the least amount of storage and enables fast data processing [23].

The process of collecting data takes three months to collect in Fouh area. Fouh area was divided into four sections: Ibn-Sina, Al-Masaken, Al- Inshaat, and Al-mutadrren area. Data was collected through direct interviews with property owners, tenants and real estate offices. Recording the data and the geographical location of the property on paper at the same moment and after collecting a sufficient number of data, the data was placed in an excel file, and then the necessary operations were performed in the data file in order to be data that can be used in creating a model.

The main reason for collecting data is that real estate data is environmental, and there is no data specific to the target region, although there are many global data available for real estate for other regions.

3.3 Data Description

Through field visits to real estate realtors and real estate owners who were interviewed and consulted, 33 attributes were identified and the focus was on these attributes because they have an impact on the real estate estimated rental of the property through which will get the prediction, and some attributes that may have an effect on the prediction process have been ignored, such as the color of the walls, escalators, heaters, etc., because they are not available or difficult to obtain, The following points describe the attributes on which the data was collected.

1. Coordinates

In this attribute, the geographical location of the property is specified, and the data type for this attribute is considered to be Geographic.

2. Property-Type

In this attribute, the type of property is determined if it is an apartment, store or other, and the data type for this attribute is categorical. In the collected data, only 540 apartments and 111 store were obtained, and the total number reached 651 store and apartments, as shown in the following figure.

| الحي | | | | |
|------------|----------|-----------|-----------|---------|
| نوع العقار | ابن سينا | الانشاءات | المتضررين | المساكن |
| شقة | 63 | 208 | 157 | 112 |
| محل | 63 | | | 48 |

Table 3.1 Number and locations of data collected

3. Country

In this attribute, the country in which the property is located is specified. The data that was collected are all in the Republic of Yemen only. The data for this attribute is of type String.

4. Governorate

In this attribute, the governorate in which the property is located is specified. The data that was collected are all in Hadramout Governorate only. The data for this attribute is of a type String.

5. City

In this attribute, the city in which the property is located is specified. The data that was collected are all in Mukalla city only. The data for this attribute is of type string.

6. Neighborhood

In this attribute, the neighborhood in which the property is located is specified. The data collected is located in four neighborhoods, and they are Al-Inshat, Al-Motdarren, Al-Masakin, and Ibn Sina. The data for this attribute is of a string type.

7. Mountain/Marine

In this attribute, it is determined whether the location of the property is located on the mountainous side relative to the main street or the marine side. There are 395 record was collected on the mountain side and 256 in the Marine side. The data type for this attribute is of the categorical type, and the following figure shows the statistics.

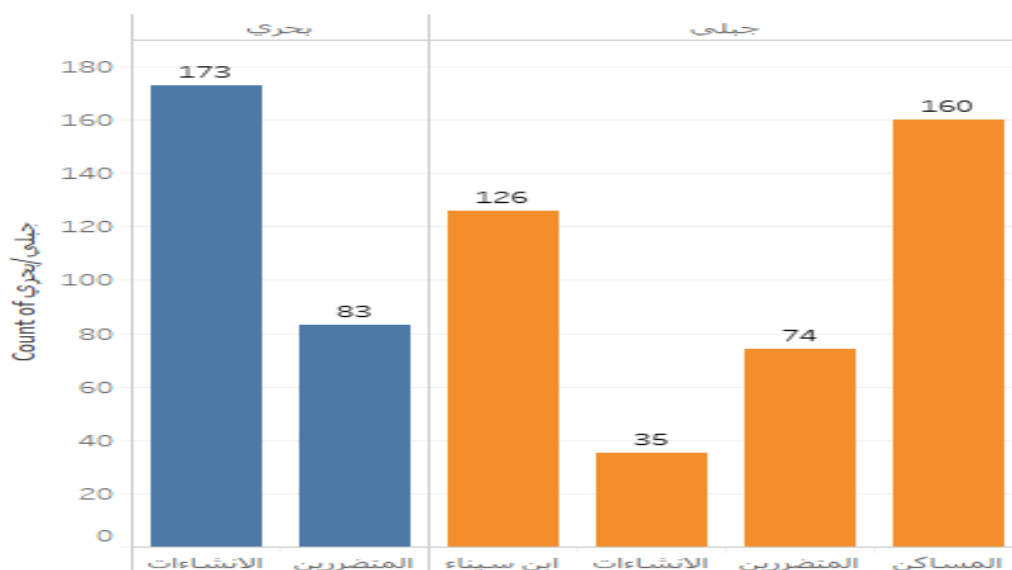


Figure 3.2 Data collected from marine side and mountain side.

8. Street-Type

In this attribute, it is determined whether the street opposite to the property is considered a main or secondary street, The data type for this attribute is categorical.

9. Adjacent side

In this attribute, the number of sides adjacent to the property are determined, the data type for this attribute is numerical.

10. Side

In this attribute, the side opposite the property is determined, whether it was the eastern, western, southern or north side, the data type for this attribute is categorical.

11. Price

In this attribute, the value of the property is determined, relying on the Saudi riyal to determine the value of the property, and the data type for this attribute is numerical.

12. N-rooms

In this attribute, the number of rooms in the property is determined, the data type for this attribute is numerical.

13. N-bathrooms

In this attribute, the number of bathrooms in the property is determined, the data type for this attribute is numerical.

14. N-kitchens

In this attribute, the number of kitchens in the property is determined, the data type for this attribute is numerical.

15. N-halls

In this attribute, the number of halls in the property is determined, the data type for this attribute is numerical.

16. Residential /Commercial

In this attribute, it is determined if the property is available for commercial or residential or both, and the type of data for this attribute categorical, and the following figure shows a statistic for the number of residential and commercial properties.

| | الحي | | | |
|--------------|---------|-----------|-----------|-----------|
| | المساكن | المتضررين | الانشاءات | ابن سيناء |
| سكني تجاري | 60 | 35 | 2 | 64 |
| سكني | 94 | 122 | 198 | 62 |
| سكني و تجاري | 6 | | 8 | |

Table 3.2 Residential/Commercial Real Estate

17. N-floor

In this attribute, the floor number of the property is determined, where the floor number is one of the most important criteria that determine the real estate estimated rental, the data type for this numerical attribute.

18. Area

In this attribute, the area of the property in length and width is specified, and the type of this attribute is Numerical.

19. Furniture/non -Furnished

In this attribute, it is specified if the property is furnished or unfurnished, and the data type of this attribute is categorical.

20. Deluxe/ Standard

In this attribute, it is determined if the property is deluxe or standard, this attribute will help us determine the quality and luxury of the property, and the data type of this attribute categorical and the following figure shows the statistics of this attribute.

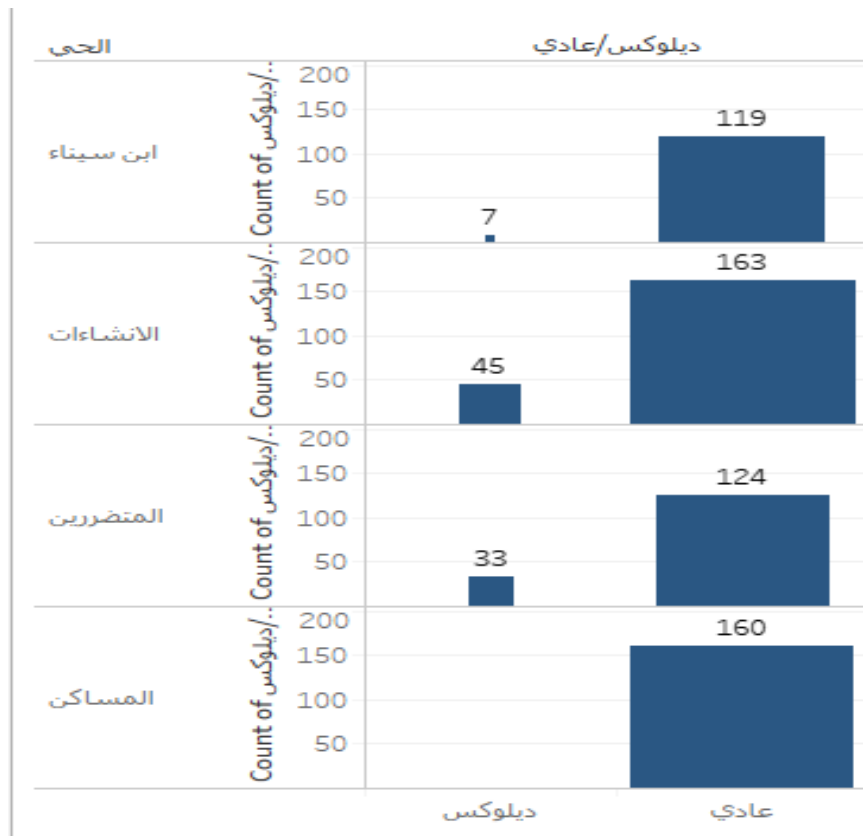


Figure 3.4 number of the deluxe and standard property.

21. Contract

In this attribute, the type of the lease contract is determined, whether it is semi-annual, annual or other, and the data type of this attribute String, and the following figure shows statistics for the type of lease.

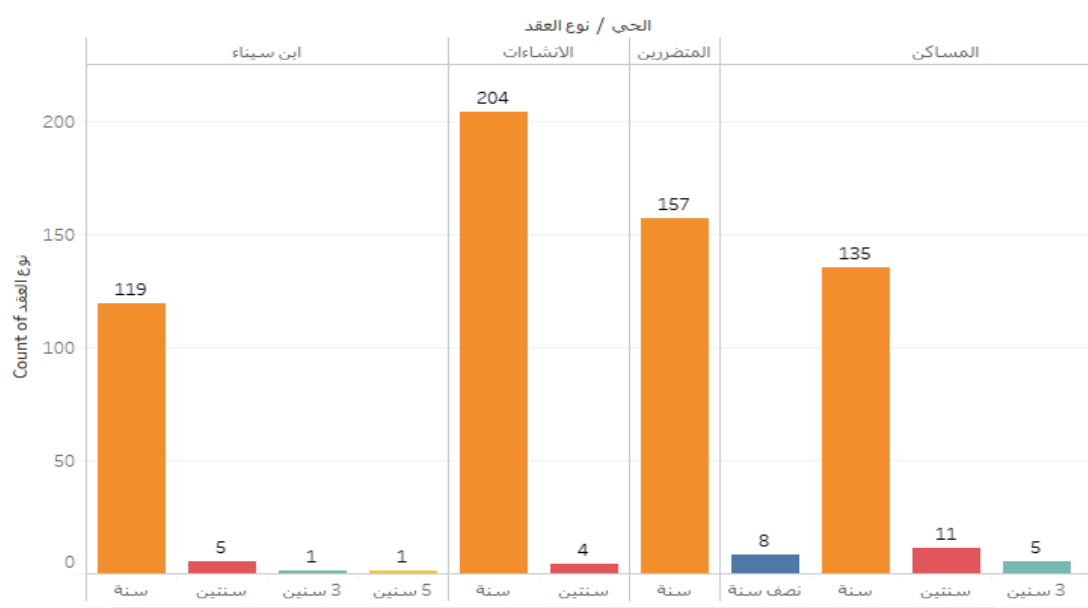


Figure 3.5 Describe the types of real estate contracts collected.

22. Concrete/Non-Concrete

In this attribute, it is determined if the property is concrete or non-concrete. This attribute is important and is useful for determining the strength of the property, and the type of data for this attribute is categorical, and the following figure illustrates.



Figure 3.6 Represent the number of concrete and non-concrete property.

23. N-balconies

In this attribute, the number of balconies in the property is determined, the data type for this feature is Numerical.

24. Families/ singles

In this attribute, it is determined whether renting the property is allowed for singles or families, or both, and the type of data for this attribute is categorical.

25. Property Age

In this attribute, the age of the property is determined whether it is new or old. The classification has been approved on the basis that the property over 16 years old is considered old and 16 and less than 16 years old is considered new, and the data type of this attribute is string.

The equation used to classify the property age attribute is show in below:

$$\frac{\sum_{i=0}^n(\text{property ages})}{(\text{Number of property ages})}$$

Equation 3.1 property age

26. Services

In this attribute, it is specified if the property has an excellent, good or bad service, examples of services are the water and the electricity, and the data type of this attribute is categorical.

27. Population-density

In this attribute, it is specified if the property has a high, medium or low population density, and the data type of this attribute is categorical.

28. Rental-period

In this attribute, the period in which the property is rented is determined, whether it is newly rented, old or not rented yet. The classifications have been approved those 5 years and less is considered a modern tenant and more than five years is an old tenant and the data type of this attribute is categorical.

The equation used to classify the rental period attribute is show in below:

$$\frac{\sum_{i=0}^n(\text{Rental periods})}{(\text{Number of Rental periods})}$$

Equation 3.2 Rental period

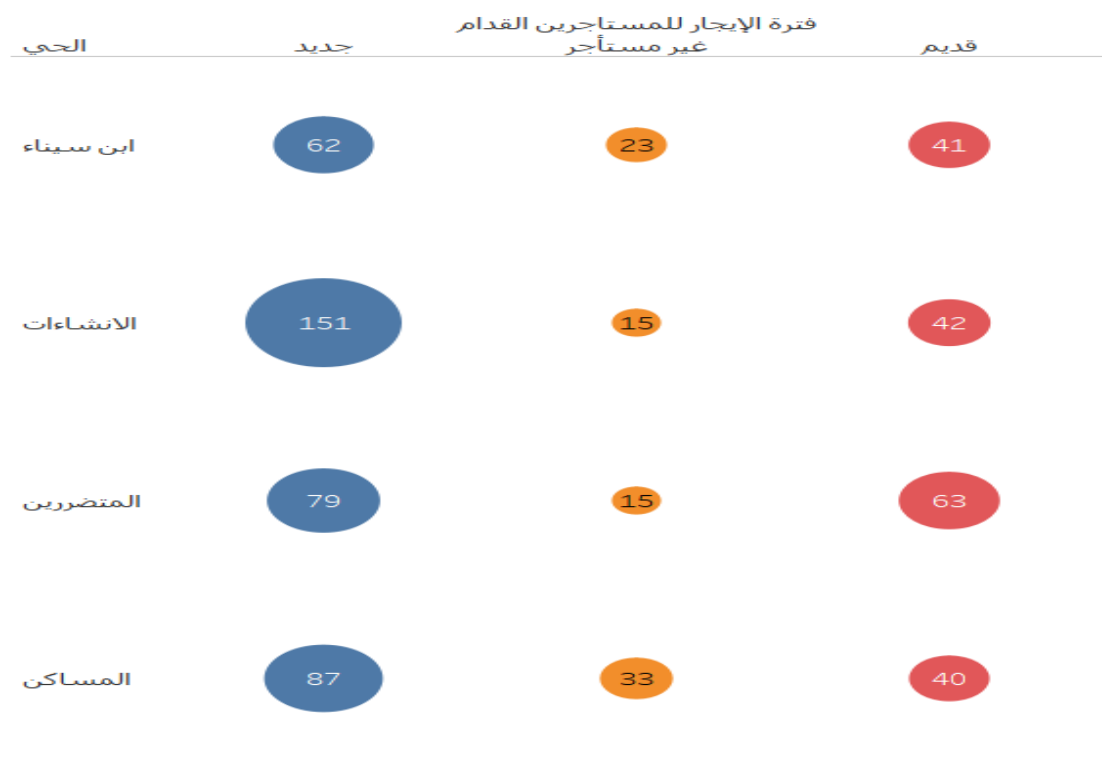


Figure 3.7 Description of the rental period for old tenants for each area.

29. Water-meter

In this feature, it is specified whether the water meter is private or shared for all tenants, and the data type of this attribute is categorical.

30. N-entrances

In this attribute, the number of property entries is specified, and the data type of this attribute is Numerical.

31. Access-road

In this attribute, the type of road leading to the property was specified, whether it was paved or landfilled, and the data type of this attribute was categorical.

Sheet 1

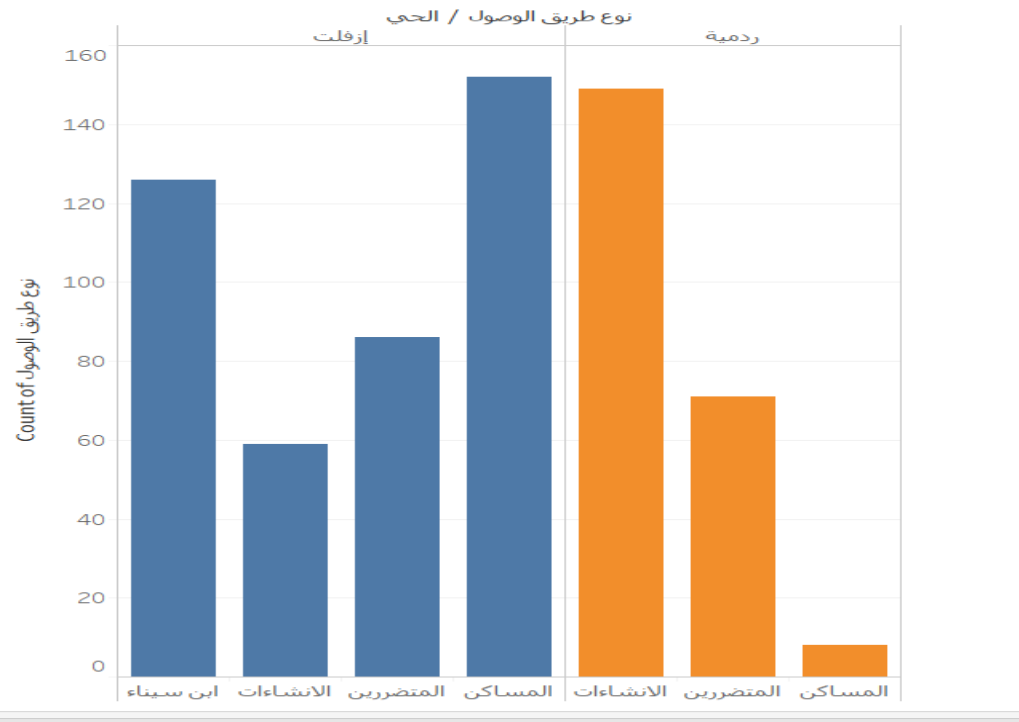


Figure 3.8 Access road type stats for each region.

32. Distance

In this attribute, the distance between the specified property and the main street is specified, and the data type for this attribute is Numerical.

33. Property Condition

In this attribute, it is determined whether the condition of the property is rented or not. The number of 561 rented properties and 90 non-rented properties have been collected, and the type of this attribute is categorical.

| الحى | الحالة | |
|-----------|--------|------------|
| | مستأجر | غير مستأجر |
| ابن سينا | 102 | 24 |
| الانشاءات | 193 | 15 |
| المتضررين | 142 | 15 |
| المساكن | 124 | 36 |

Table 3.3 Property condition for each area.

3.4 Data preparation & cleaning

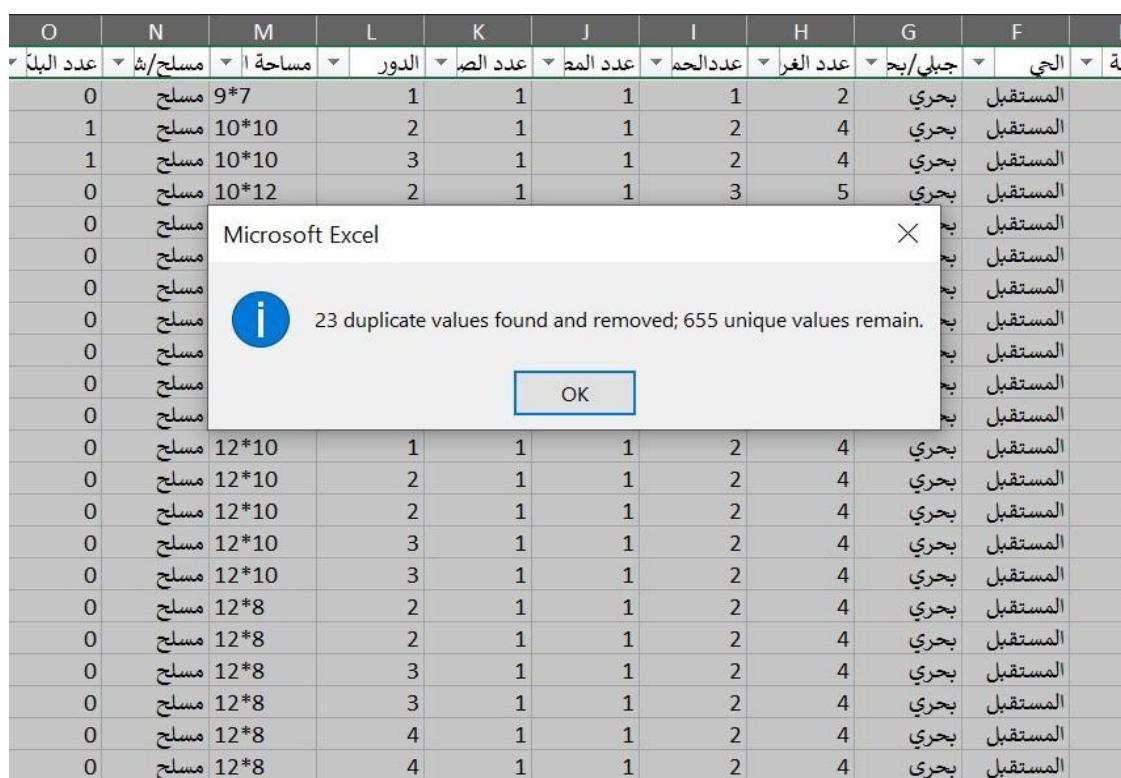
Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data [24].

Data preparation is often a lengthy undertaking for data professionals or business users, but it is essential as a prerequisite to put data in context in order to turn it into insights and eliminate bias resulting from poor data quality [24]. The process of cleaning the data is shown in the following steps.

3.4.1 Remove Duplicate

At this stage, duplicate data is removed, and often duplicate data occurs during data collection.

In the collected data, 23 duplicate records were found and they were removed using Excel. The following figure illustrates this operation.



| O | N | M | L | K | J | I | H | G | F | E |
|-----------|----------|---------|-------|----------|----------|----------|----------|---------|----------|-----|
| عدد البلد | مساحه/نم | مساحه ا | الدور | عدد الصم | عدد المم | عدد الحم | عدد الغر | جبلي/بح | الحي | بنه |
| 0 | مساح | 9*7 | 1 | 1 | 1 | 1 | 2 | بحري | المستقبل | |
| 1 | مساح | 10*10 | 2 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 1 | مساح | 10*10 | 3 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 0 | مساح | 10*12 | 2 | 1 | 1 | 3 | 5 | بحري | المستقبل | |
| 0 | مساح | | | | | | | | المستقبل | |
| 0 | مساح | | | | | | | | المستقبل | |
| 0 | مساح | | | | | | | | المستقبل | |
| 0 | مساح | | | | | | | | المستقبل | |
| 0 | مساح | | | | | | | | المستقبل | |
| 0 | مساح | | | | | | | | المستقبل | |
| 0 | مساح | | | | | | | | المستقبل | |
| 0 | مساح | | | | | | | | المستقبل | |
| 0 | مساح | | | | | | | | المستقبل | |
| 0 | مساح | 12*10 | 1 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 0 | مساح | 12*10 | 2 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 0 | مساح | 12*10 | 2 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 0 | مساح | 12*10 | 3 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 0 | مساح | 12*10 | 3 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 0 | مساح | 12*8 | 2 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 0 | مساح | 12*8 | 2 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 0 | مساح | 12*8 | 3 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 0 | مساح | 12*8 | 3 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 0 | مساح | 12*8 | 4 | 1 | 1 | 2 | 4 | بحري | المستقبل | |
| 0 | مساح | 12*8 | 4 | 1 | 1 | 2 | 4 | بحري | المستقبل | |

Figure 3.10 delete duplicates by excel

3.4.2 Handel Missing Values

Missing data is common in many different areas of data science and machine learning. Unfortunately, it can be challenging to handle effectively, and often there is no best solution [25].

Missing data cannot be ignored because many algorithms do not accept missing values and there are two ways to deal with missing data.

The first method is to deal with missing data by filling it in. In this method, it is required that the missing data be predictable and that the filling be logical and based on probabilities and assumptions.

The missing data was found in the area column and was filled in by looking at the price of the apartment and the number of its rooms, and looking at the previous data as shown in the following figures.

| R | Q | P | O | N | M | L | K | J | I | H | G | F | E | D | C | B | A | |
|---|------------|------|---|------|-------|---|---|---|---|---|------|-----------|--------|--------|-------|-----|-------------------------|-----|
| 1 | جديد | قديم | 2 | مساح | 10*10 | 2 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرهوت | اليمن | شقة | 14.5008180, 49.0576760 | 494 |
| 1 | جديد | قديم | 2 | مساح | 10*10 | 2 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرهوت | اليمن | شقة | 14.5008180, 49.0576760 | 495 |
| 1 | جديد | قديم | 2 | مساح | 10*10 | 2 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرهوت | اليمن | شقة | 14.5008180, 49.0576760 | 496 |
| 1 | جديد | قديم | 2 | مساح | 10*10 | 3 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرهوت | اليمن | شقة | 14.5008180, 49.0576760 | 497 |
| 1 | جديد | قديم | 2 | مساح | 10*10 | 3 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرهوت | اليمن | شقة | 14.5008180, 49.0576760 | 498 |
| 1 | جديد | قديم | 2 | مساح | 10*10 | 3 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرهوت | اليمن | شقة | 14.5008180, 49.0576760 | 499 |
| 1 | جديد | قديم | 2 | مساح | 10*10 | 3 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرهوت | اليمن | شقة | 14.5008180, 49.0576760 | 500 |
| 1 | جديد | قديم | 0 | مساح | | 1 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرهوت | اليمن | شقة | 556, 14.471727914961576 | 501 |
| 1 | جديد | قديم | 1 | مساح | | 2 | 1 | 1 | 2 | 5 | جيلي | ابن سيناء | المكلا | حضرهوت | اليمن | شقة | 556, 14.471727914961576 | 502 |
| 1 | جديد | قديم | 1 | مساح | | 3 | 1 | 1 | 2 | 5 | جيلي | ابن سيناء | المكلا | حضرهوت | اليمن | شقة | 556, 14.471727914961576 | 503 |
| 1 | جديد | قديم | 1 | مساح | | 4 | 1 | 1 | 1 | 2 | جيلي | ابن سيناء | المكلا | حضرهوت | اليمن | شقة | 556, 14.471727914961576 | 504 |
| 1 | جديد | قديم | 0 | مساح | | 1 | 1 | 1 | 1 | 3 | جيلي | ابن سيناء | المكلا | حضرهوت | اليمن | شقة | 502, 14.470539015775778 | 505 |
| 1 | جديد | قديم | 1 | مساح | | 2 | 1 | 1 | 2 | 5 | جيلي | ابن سيناء | المكلا | حضرهوت | اليمن | شقة | 502, 14.470539015775778 | 506 |
| 1 | جديد | قديم | 2 | مساح | | 2 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرهوت | اليمن | شقة | 14.471873, 49.029960 | 507 |
| 1 | جديد | قديم | 2 | مساح | | 3 | 2 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرهوت | اليمن | شقة | 14.471873, 49.029960 | 508 |
| 1 | جديد | قديم | 2 | مساح | | 4 | 3 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرهوت | اليمن | شقة | 14.471873, 49.029960 | 509 |
| 1 | جديد | قديم | 2 | مساح | | 3 | 1 | 1 | 2 | 5 | جيلي | ابن سيناء | المكلا | حضرهوت | اليمن | شقة | 399, 14.468560358286847 | 510 |
| 1 | جديد | قديم | 2 | مساح | | 2 | 1 | 1 | 2 | 6 | جيلي | ابن سيناء | المكلا | حضرهوت | اليمن | شقة | 399, 14.468560358286847 | 511 |
| 1 | غير مستأجر | جديد | 4 | مساح | 20*20 | 2 | 1 | 1 | 3 | 5 | بحري | الاشاعات | المكلا | حضرهوت | اليمن | شقة | 14.5063357, 49.0614565 | 512 |
| 1 | غير مستأجر | جديد | 4 | مساح | 20*20 | 3 | 1 | 1 | 3 | 5 | بحري | الاشاعات | المكلا | حضرهوت | اليمن | شقة | 14.5063357, 49.0614565 | 513 |
| 1 | جديد | قديم | 1 | مساح | 8*16 | 1 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرهوت | اليمن | شقة | 14.5038442, 49.0545214 | 514 |
| 1 | جديد | جديد | 1 | مساح | 8*16 | 1 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرهوت | اليمن | شقة | 14.5038442, 49.0545214 | 515 |

Figure 3.11 Example of data before filling in the missing records

| R | Q | P | O | N | M | L | K | J | I | H | G | F | E | D | C | B | A | | |
|---|----------|------|---|------|-------|---|---|---|---|---|------|-----------|-----------|--------|--------|-------|------------------------|------------------------|-----|
| 1 | جديد | قديم | 2 | مسلح | 10*10 | 2 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرعوت | اليمن | شفة | 14.5008180,49.0576760 | 495 | |
| 1 | جديد | قديم | 2 | مسلح | 10*10 | 2 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرعوت | اليمن | شفة | 14.5008180,49.0576760 | 496 | |
| 1 | جديد | قديم | 2 | مسلح | 10*10 | 3 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرعوت | اليمن | شفة | 14.5008180,49.0576760 | 497 | |
| 1 | جديد | قديم | 2 | مسلح | 10*10 | 3 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرعوت | اليمن | شفة | 14.5008180,49.0576760 | 498 | |
| 1 | جديد | قديم | 2 | مسلح | 10*10 | 3 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرعوت | اليمن | شفة | 14.5008180,49.0576760 | 499 | |
| 1 | جديد | قديم | 2 | مسلح | 10*10 | 3 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرعوت | اليمن | شفة | 14.5008180,49.0576760 | 500 | |
| 1 | جديد | قديم | 0 | مسلح | 10*10 | 1 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 556,14.471727914961576 | 501 | |
| 1 | جديد | قديم | 1 | مسلح | 10*10 | 2 | 1 | 1 | 2 | 5 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 556,14.471727914961576 | 502 | |
| 1 | جديد | قديم | 1 | مسلح | 10*10 | 3 | 1 | 1 | 2 | 5 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 556,14.471727914961576 | 503 | |
| 1 | جديد | قديم | 1 | مسلح | 10*10 | 4 | 1 | 1 | 2 | 1 | 2 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 556,14.471727914961576 | 504 |
| 1 | جديد | قديم | 0 | مسلح | 8*8 | 1 | 1 | 1 | 1 | 3 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 502,14.470539015775778 | 505 | |
| 1 | جديد | قديم | 1 | مسلح | 12*12 | 2 | 1 | 1 | 2 | 5 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 502,14.470539015775778 | 506 | |
| 1 | جديد | قديم | 2 | مسلح | 10*10 | 2 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.471873,49.029960 | 507 | |
| 1 | جديد | قديم | 2 | مسلح | 10*10 | 3 | 2 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.471873,49.029960 | 508 | |
| 1 | جديد | قديم | 2 | مسلح | 10*10 | 4 | 3 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.471873,49.029960 | 509 | |
| 1 | جديد | قديم | 2 | مسلح | 12*12 | 3 | 1 | 1 | 2 | 5 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 399,14.468560358286847 | 510 | |
| 1 | جديد | قديم | 2 | مسلح | 12*12 | 2 | 1 | 1 | 2 | 5 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 399,14.468560358286847 | 511 | |
| 1 | غير مسجل | جديد | 4 | مسلح | 20*20 | 2 | 1 | 1 | 3 | 5 | بحري | الاشاعات | المكلا | حضرعوت | اليمن | شفة | 14.5063357,49.0614565 | 512 | |
| 1 | غير مسجل | جديد | 4 | مسلح | 20*20 | 3 | 1 | 1 | 3 | 5 | بحري | الاشاعات | المكلا | حضرعوت | اليمن | شفة | 14.5063357,49.0614565 | 513 | |
| 1 | جديد | جديد | 1 | مسلح | 8*16 | 1 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرعوت | اليمن | شفة | 14.5038442,49.0545214 | 514 | |
| 1 | جديد | جديد | 1 | مسلح | 8*16 | 1 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرعوت | اليمن | شفة | 14.5038442,49.0545214 | 515 | |
| 1 | جديد | جديد | 1 | مسلح | 8*16 | 2 | 1 | 1 | 2 | 3 | جيلي | الاشاعات | المكلا | حضرعوت | اليمن | شفة | 14.5038442,49.0545214 | 516 | |

Figure 3.12 Example of data after filling in the missing records

The second method of dealing with lost data when it is impossible to predict the lost data is by delete the record.

In Figure (3.13) missing data was found in the geographic coordinate column, that data was deleted due to its unpredictability and in Figure (3.14) the missing data was deleted.

| R | Q | P | O | N | M | L | K | J | I | H | G | F | E | D | C | B | A | |
|------|------------|------|---|------|-------|---|---|---|---|---|------|-----------|--------|--------|-------|-----|-------------------------|-----|
| سنوي | غير مستاجر | قديم | 2 | مسلح | 12*12 | 2 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4696585,49.0299413 | 355 |
| سنوي | غير مستاجر | قديم | 2 | مسلح | 12*12 | 3 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4696585,49.0299413 | 356 |
| سنوي | غير مستاجر | قديم | 0 | مسلح | 24*12 | 1 | 2 | 2 | 4 | 8 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4701494,49.0292881 | 357 |
| سنين | قديم | قديم | 0 | مسلح | 12*12 | 1 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.469549943704939,49.0 | 358 |
| سنين | قديم | قديم | 0 | مسلح | 12*12 | 2 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.469549943704939,49.0 | 359 |
| سنين | قديم | قديم | 0 | مسلح | 12*12 | 3 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.469549943704939,49.0 | 360 |
| سنين | قديم | قديم | 0 | مسلح | 6*6 | 4 | 1 | 1 | 1 | 1 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.469549943704939,49.0 | 361 |
| سنوي | قديم | قديم | 0 | مسلح | 12*12 | 1 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4696952,49.0296616 | 362 |
| سنوي | قديم | قديم | 0 | مسلح | 12*12 | 2 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4696952,49.0296616 | 363 |
| سنوي | قديم | قديم | 0 | مسلح | 12*12 | 3 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4696952,49.0296616 | 364 |
| سنوي | غير مستاجر | قديم | 0 | مسلح | 6*6 | 4 | 1 | 1 | 1 | 1 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4696952,49.0296616 | 365 |
| سنوي | غير مستاجر | قديم | 0 | مسلح | 12*12 | 1 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | | 366 |
| سنوي | قديم | قديم | 2 | مسلح | 12*12 | 2 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | | 367 |
| سنوي | قديم | قديم | 0 | مسلح | 10*10 | 1 | 1 | 1 | 2 | 5 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4701477,49.0296563 | 368 |
| سنوي | قديم | قديم | 0 | مسلح | 10*10 | 1 | 1 | 1 | 2 | 5 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4701477,49.0296563 | 369 |
| سنوي | قديم | قديم | 0 | مسلح | 8*12 | 1 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4696543,49.0292731 | 370 |
| سنوي | قديم | قديم | 2 | مسلح | 12*12 | 2 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4696543,49.0292731 | 371 |
| سنوي | غير مستاجر | قديم | 0 | مسلح | 10*10 | 1 | 1 | 1 | 2 | 3 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4704863,49.0294605 | 372 |
| سنوي | غير مستاجر | قديم | 2 | مسلح | 12*12 | 2 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4704863,49.0294605 | 373 |
| سنوي | غير مستاجر | قديم | 2 | مسلح | 12*12 | 3 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4704863,49.0294605 | 374 |
| سنوي | غير مستاجر | قديم | 2 | مسلح | 12*12 | 4 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4704863,49.0294605 | 375 |
| سنوي | غير مستاجر | قديم | 2 | مسلح | 12*12 | 5 | 1 | 1 | 2 | 4 | جيلي | ابن سيناء | المكلا | حضرعوت | اليمن | شفة | 14.4704863,49.0294605 | 376 |

Figure 3.13 Example of data before deleting lost records.

| R | Q | P | O | N | M | L | K | J | I | H | G | F | E | D | C | B | A | |
|-----------------|------|---|------|-------|---|---|---|---|---|-------|-----------|--------|---------|------|-----|-----------------------------|-----|--|
| غير مستاجر سنوي | قديم | 2 | شعبي | 12*12 | 4 | 1 | 2 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4710171,49.0293073 | 353 | |
| غير مستاجر سنوي | قديم | 2 | شعبي | 12*12 | 5 | 1 | 2 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4710171,49.0293073 | 354 | |
| غير مستاجر سنوي | قديم | 2 | مساح | 12*12 | 2 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4696585,49.0299413 | 355 | |
| غير مستاجر سنوي | قديم | 2 | مساح | 12*12 | 3 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4696585,49.0299413 | 356 | |
| غير مستاجر سنوي | جديد | 0 | مساح | 24*12 | 1 | 2 | 2 | 4 | 8 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4701494,49.0292881 | 357 | |
| سنتين | قديم | 0 | مساح | 12*12 | 1 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.469549943704939, 49.0358 | 358 | |
| سنتين | جديد | 0 | مساح | 12*12 | 2 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.469549943704939, 49.0359 | 359 | |
| سنتين | جديد | 0 | مساح | 12*12 | 3 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.469549943704939, 49.0360 | 360 | |
| سنتين | جديد | 0 | مساح | 6*6 | 4 | 1 | 1 | 1 | 1 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.469549943704939, 49.0361 | 361 | |
| سنوي | جديد | 0 | مساح | 12*12 | 1 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4696952,49.0296616 | 362 | |
| سنوي | جديد | 0 | مساح | 12*12 | 2 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4696952,49.0296616 | 363 | |
| سنوي | جديد | 0 | مساح | 12*12 | 3 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4696952,49.0296616 | 364 | |
| غير مستاجر سنوي | جديد | 0 | مساح | 6*6 | 4 | 1 | 1 | 1 | 1 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4696952,49.0296616 | 365 | |
| سنوي | جديد | 0 | مساح | 10*10 | 1 | 1 | 1 | 2 | 5 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4701477,49.0296563 | 366 | |
| سنوي | جديد | 0 | مساح | 10*10 | 1 | 1 | 1 | 2 | 5 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4701477,49.0296563 | 367 | |
| سنوي | جديد | 0 | مساح | 8*12 | 1 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4696543,49.0292731 | 368 | |
| سنوي | جديد | 2 | مساح | 12*12 | 2 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4696543,49.0292731 | 369 | |
| غير مستاجر سنوي | جديد | 0 | مساح | 10*10 | 1 | 1 | 1 | 2 | 3 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4704863,49.0294605 | 370 | |
| غير مستاجر سنوي | جديد | 2 | مساح | 12*12 | 2 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4704863,49.0294605 | 371 | |
| غير مستاجر سنوي | جديد | 2 | مساح | 12*12 | 3 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4704863,49.0294605 | 372 | |
| غير مستاجر سنوي | جديد | 2 | مساح | 12*12 | 4 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4704863,49.0294605 | 373 | |
| غير مستاجر سنوي | جديد | 2 | مساح | 12*12 | 5 | 1 | 1 | 2 | 4 | جبلبي | ابن سنياء | المكلا | حضر موت | البن | شقة | 14.4704863,49.0294605 | 374 | |

Figure 3.14 Example of data after deleting lost records.

3.4.3 Filter outliers

Outliers data is the data does not fit with the data being analyzed and may cause a problem in performing the purpose for which the data was collected.

There are a lot of methods to detect these outliers, the interquartile range method is one of them which used to find the upper and lower bounders, and any value out of these bounders is outlier value, and the following interquartile range equation is the used equation:

$$IQR = Q_3 - Q_1$$

Equation 3.3 Interquartile range [27]

In this equation, a typical value is summarizing using the median as opposed to the mean, which is the difference between the first and third quartiles.

The lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order. The following figure shows an example of the detection of the outlier values using the interquartile range equation, where the upper bounder is equal 1300 and the lower bounder is equal 100 and all the data inside the gray rectangle is on the range otherwise, they are outliers [27].

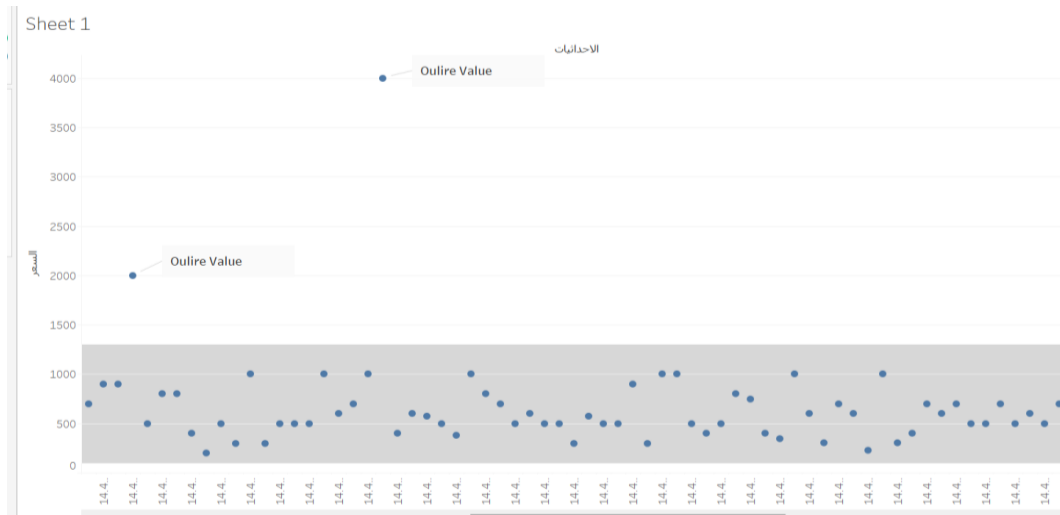


Figure 3.15 Clarification of data before deleting outliers

After detecting the outliers, they have been dealt with by removing it from the data.

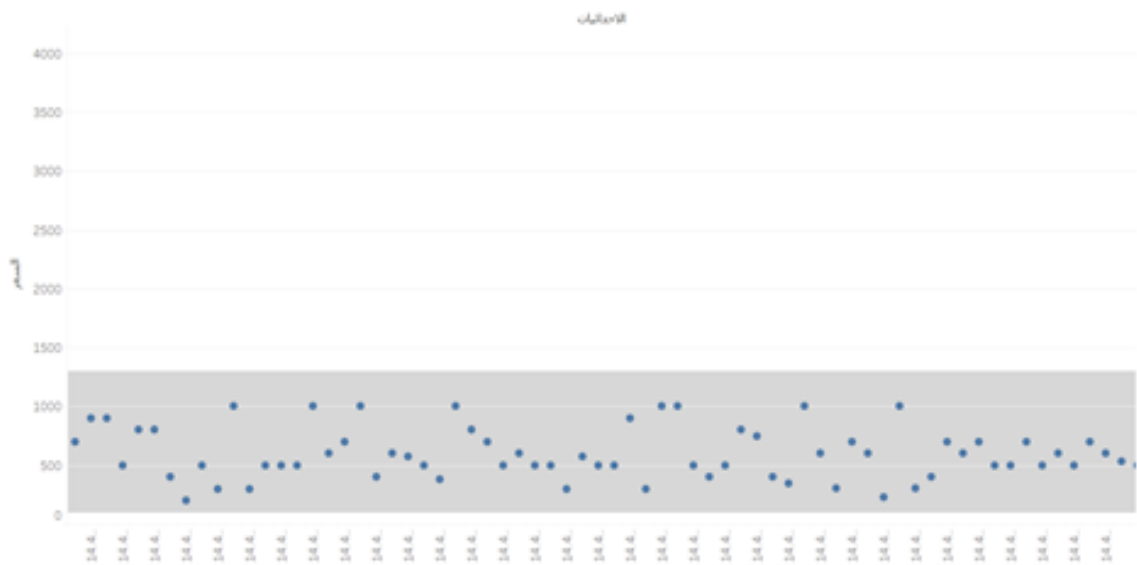


Figure 3.16 Clarification of data after deleting outliers

3.5 Final prepared data

After defining the requirements and collecting and cleaning the data, 591 records of apartments and shops distributed in four regions have obtained, with 82 stores and 509 apartments. 192 apartments in the Al-Inshaat area on its marine and mountain sides have obtained, including 35 mountain apartments and 157 marine apartments, and 156 apartments in the Al-mutadren area on its mountainous and marine sides, including 74 mountain apartments and 83 marine apartments. In the Al-Masaken

area, 110 apartments and 42 stores on the mountainous side have obtained, and in the Ibn-Sina area on its mountainous side 51 apartments and 40 stores have obtained.

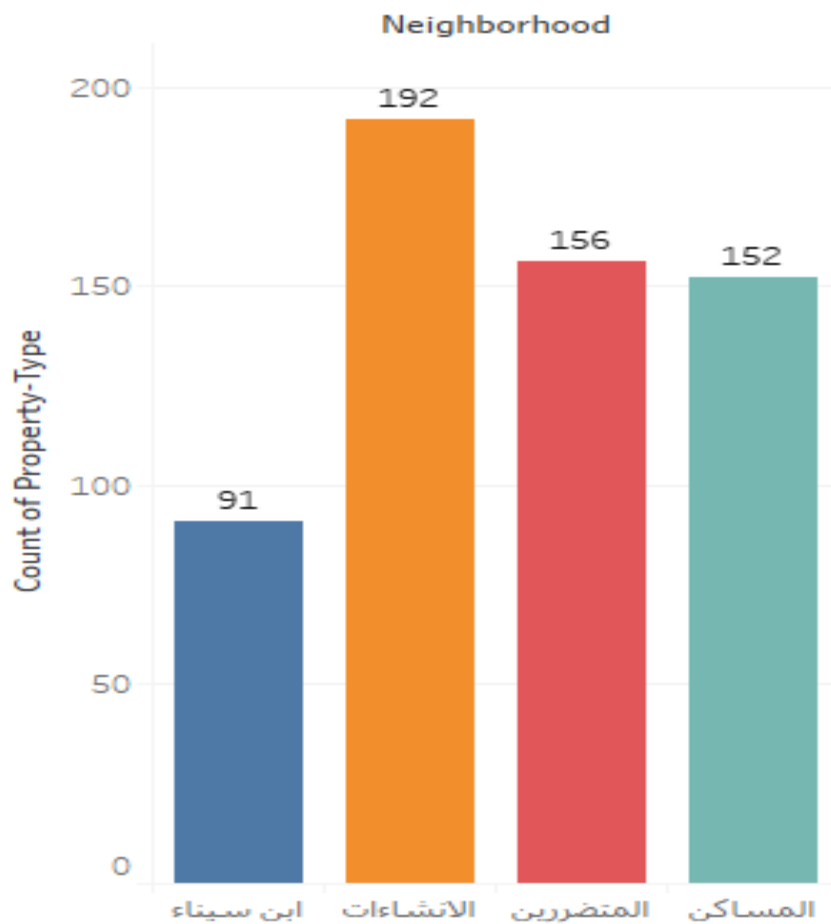


Figure 3.17 apartments after cleaning

Chapter Four

Exploratory Data Analysis

4.1 Introduction

In this chapter the data are explored starting with data descriptive statistics, then the data distribution and visualization then ending with finding the correlations.

Before getting start an introduction of the exploratory data analysis will be taken.

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It uses to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations [28].

EDA is the process of using numerical summaries and visualizations to explore the data and to identify potential relationships between variables [28].

EDA is an investigative process that uses summary statistics and graphical tools to get to know the data and understand what can be learn from them [28].

With EDA, anomalies in the data are easily found, such as outliers or unusual observations, uncover patterns, understand potential relationships among variables, and generate interesting questions or hypotheses that test later using more formal statistical methods [28].

EDA is like detective work: starts with searching for clues and insights that can lead to the identification of potential root causes of the problem need to solve. With exploring one variable at a time, then two variables at a time, and then many variables at a time [28].

Although EDA encompasses tables of summary statistics such as the mean and standard deviation, most people focus on graphs. With using of a variety of graphs and exploratory tools, and go where the data go. If one graph or analysis is not informative, look at the data from another perspective [28].

4.2 Descriptive statistics

Descriptive statistics are used to summaries and describe a variable or variables for a sample of data (as opposed to drawing conclusions about any larger population from which the sample was drawn- this is covered in the inferential statistics page). For example, sample statistics such as the mean (\bar{x}) and standard deviation (STD) are often used to summaries and describe continuous variables [29].

Furthermore, descriptive statistics can be used to summaries just one variable at a time, to analysis relationships between two variables and to analysis relationships between three or more variables [29].

Jupyter will be used in the data Descriptive statistics process.

To start the data Descriptive Statistics Process, start with calling the necessary libraries and importing the data to be works on and display it to take an overview and get an initial impression.

In [36]:

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
from scipy.stats import f_oneway
Data=pd.read_csv('RealEstate_Dataset.CSV')
Data
```

Out[36]:

| | Coordinates | Property-Type | Country | Governorate | City | Neighborhood | Mountain/Marine | N-rooms | N-bathrooms | N-kitchens | ... | Water-meter | Access-road | Popl |
|-----|-----------------------|---------------|---------|-------------|--------|--------------|-----------------|---------|-------------|------------|-----|-------------|-------------|------|
| 0 | 14.4923994,49.0553294 | شقة | اليمن | حضرموت | المكلا | المقصرين | بحري | 4 | 2 | 1 | ... | خاص | رسمية | |
| 1 | 14.4927351,49.0553579 | شقة | اليمن | حضرموت | المكلا | المقصرين | بحري | 4 | 2 | 1 | ... | خاص | رسمية | |
| 2 | 14.4927351,49.0553579 | شقة | اليمن | حضرموت | المكلا | المقصرين | بحري | 5 | 3 | 1 | ... | خاص | رسمية | |
| 3 | 14.4927351,49.0553579 | شقة | اليمن | حضرموت | المكلا | المقصرين | بحري | 5 | 3 | 1 | ... | خاص | رسمية | |
| 4 | 14.4927351,49.0553579 | شقة | اليمن | حضرموت | المكلا | المقصرين | بحري | 5 | 3 | 1 | ... | خاص | رسمية | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 586 | 14.4880623,49.0444741 | محل | اليمن | حضرموت | المكلا | المساكن | جبلي | 1 | 0 | 0 | ... | خاص | إرقلت | |
| 587 | 14.4880990,49.0444886 | محل | اليمن | حضرموت | المكلا | المساكن | جبلي | 1 | 0 | 0 | ... | خاص | إرقلت | |
| 588 | 14.4881146,49.0445090 | محل | اليمن | حضرموت | المكلا | المساكن | جبلي | 1 | 0 | 0 | ... | خاص | إرقلت | |
| 589 | 14.4881857,49.0445667 | محل | اليمن | حضرموت | المكلا | المساكن | جبلي | 2 | 0 | 0 | ... | خاص | إرقلت | |
| 590 | 14.4883077,49.0447695 | محل | اليمن | حضرموت | المكلا | المساكن | جبلي | 1 | 0 | 0 | ... | خاص | إرقلت | |

Figure 4.1: Snapshot of the dataset

The shape function is used to show the number of records and columns.

In [37]:

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
Data.shape
```

Out[37]:

(591, 33)

There are 33 columns and 591 records.

Here the info () function has used to show the type of data and what columns are likely to have missing values, in addition to the number of columns and records.

```
In [4]: 1 Data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 591 entries, 0 to 590
Data columns (total 33 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Coordinates                           591 non-null    object
1   Property-Type                         591 non-null    object
2   Country                              591 non-null    object
3   Governorate                          591 non-null    object
4   City                                 591 non-null    object
5   Neighborhood                         591 non-null    object
6   Mountain/Marine                      591 non-null    object
7   N-rooms                             591 non-null    int64
8   N-bathrooms                         591 non-null    int64
9   N-kitchens                          591 non-null    int64
10  N-halls                             591 non-null    int64
11  N-floor                             591 non-null    int64
12  N-balconies                         591 non-null    int64
13  Property Age                         591 non-null    object
14  Rental-period                       591 non-null    object
15  Contract                            591 non-null    int64
16  Property Condition                  591 non-null    object
17  Price                              591 non-null    int64
18  Side                               591 non-null    object
19  Residential /Commercial             591 non-null    object
20  Families/ singles                   591 non-null    object
21  Concrete/Non-Concrete               591 non-null    object
22  N-entrances                        591 non-null    int64
23  Water-meter                        591 non-null    object
24  Access-road                        591 non-null    object
25  Population-density                  591 non-null    object
26  Services                           591 non-null    object
27  Distance                           591 non-null    int64
28  Street-type                        591 non-null    object
29  Adjacent-sides                     591 non-null    int64
30  Deluxe/ Standard                   591 non-null    object
31  Furniture/non -Furnished            591 non-null    object
32  Area                               591 non-null    object
dtypes: int64(11), object(22)
memory usage: 152.5+ KB
```

Figure 4.2: Data Information.

There is no missing data, and the number of records has reached 591 records and the number of columns are 33, and there are 22 data types of the objective type and 11 of the numerical data type have obtained.

Here the describe () function is used to show the count, mean, STD, min, max, 25%, 50% and, 75% information.

In [38]:

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
Data[['Price','Distance']].describe()
```

Out[38]:

| | Price | Distance |
|-------|-------------|------------|
| count | 591.000000 | 591.000000 |
| mean | 602.142132 | 314.167513 |
| std | 193.293656 | 212.499350 |
| min | 200.000000 | 10.000000 |
| 25% | 500.000000 | 135.000000 |
| 50% | 573.000000 | 300.000000 |
| 75% | 700.000000 | 480.000000 |
| max | 1000.000000 | 950.000000 |

Figure4.3 :count, mean, STD, min, max, 25%, 50% and, 75% information for Price and Distance

The above figure, show the following information:

Count: Which represents the number of values in each column, and there are 591 records in each column.

Mean: It shows us the average values in each column, the average in the price column is 602.142, and the average in the distance column is 314.167.

STD: represents the standard deviation, which is equal to 193.293 in the price column and 212.499 in the distance column, which shows the distance of the points from the average and the extent of the dispersion of the data.

Min and Max: Which represents the minimum and maximum value in each column, the minimum value of the price column is 200, and the maximum is 1000, and in the distance column, the minimum value is 10 and the maximum value in the column is 950.

25%, 50%(median), 75%: In the first quarter, which represents the median value of the first 25% of the data, it is 500 for the price column and 135 for the distance columns, for the second quarter, whose median value is 50%, the value in the middle of the data appears, which is 573 in the price column and 300 in the distance column. And the value of the third quartile, which represents the median value of 75% of the data and is equal to 700 for the price column and 480 for the distance column.

The mode () function used to show the most frequent value in each column of the data.

In [39]:

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
Data.mode()
```

Out[39]:

| Mountain/Marine | N- rooms | N- bathrooms | N- kitchens | N- halls | N- floor | N- balconies | Property Age | Rental- period | Contract | Property Condition | Price | Side | Residential /Commercial | Families/ singles | Concrete/ Con |
|-----------------|-------------|-----------------|----------------|-------------|-------------|-----------------|-----------------|-------------------|----------|-----------------------|-------|-------|----------------------------|----------------------|------------------|
| جبلي | 4 | 2 | 1 | 1 | 2 | 0 | جديد | جديد | 1 | مستأجر | 500 | مفتوح | سكني | عوائل | |

| Side | Residential /Commercial | Families/ singles | Concrete/Non- Concrete | N- entrances | Water- meter | Access- road | Population- density | Services | Distance | Street- type | Adjacent- sides | Deluxe/ Standard | Furniture/non- Furnished | Area |
|-------|----------------------------|----------------------|---------------------------|-----------------|-----------------|-----------------|------------------------|----------|----------|-----------------|--------------------|---------------------|-----------------------------|-------|
| مفتوح | سكني | عوائل | مسلح | 2 | خاص | إزفلت | A | B | 350 | فرعي | 0 | عادي | غير مفروش | Short |

Figure4.4 Most frequent value in each column of the data.

4.3 Data Distribution and Visualization

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large datasets. The term is often used interchangeably

with others, including information graphics, information visualization and statistical graphics [30].

Data distribution is a function that specifies all possible values for a variable and also quantifies the relative frequency (probability of how often they occur). Distributions are considered any population that has a scattering of data [31].

Deluxe/Standard

In this figure, the data distribution is clarified in terms of the property's status, whether it is standard or deluxe, and 506 records of standard type and 85 deluxe were obtained.

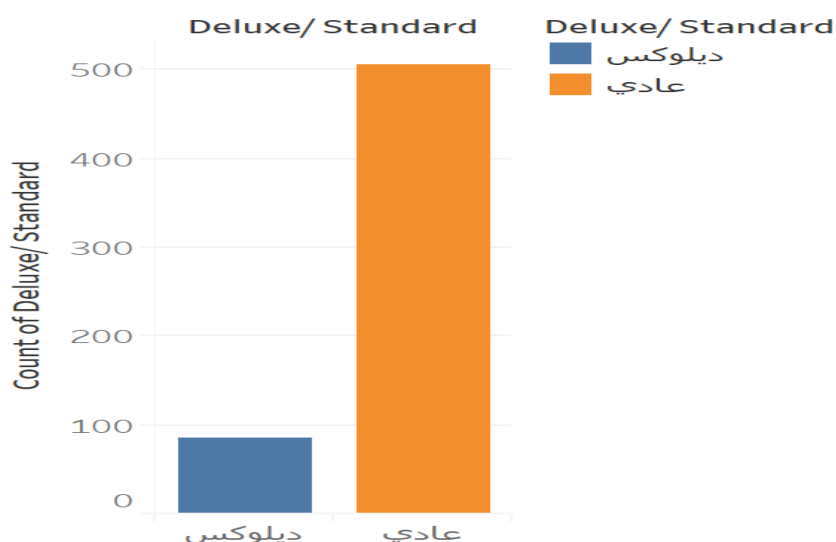


Figure4.5 data distribution is clarified in terms of the property's status, whether it is standard or deluxe

N-balconies

The following figure shows the distribution of data on categories from 0 to 4 in N-balconies column where it found 262 records containing category 0, 142 records containing category 1, 113 records containing category 2, 52 records containing category 3, and 22 records containing category 4.

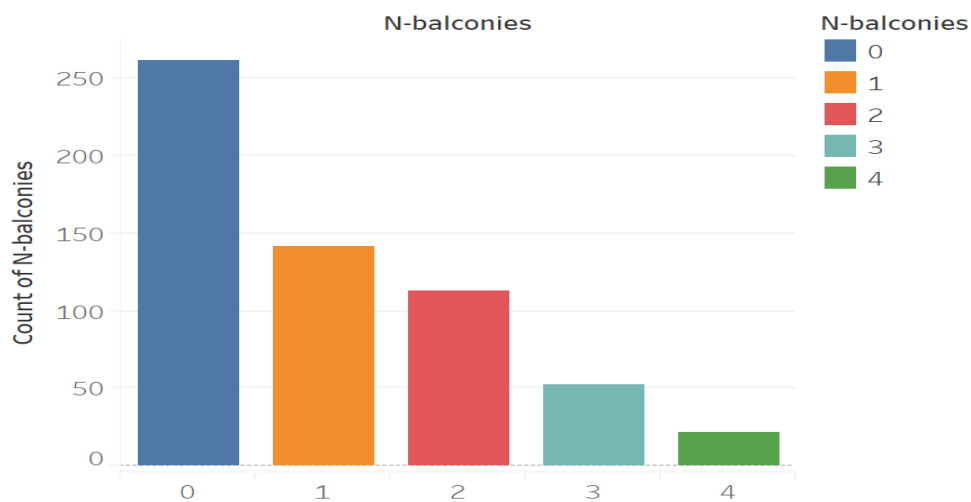


Figure 4.6 Distribution of data on categories from 0 to 4 in N-balconies column.

N-bathrooms

The following figure shows the distribution of data on categories from 0 to 3 in the N-bathrooms column, where it found 82 records containing category zero, 20 records containing category 1, 460 records containing category 2, and 29 records containing category 3.

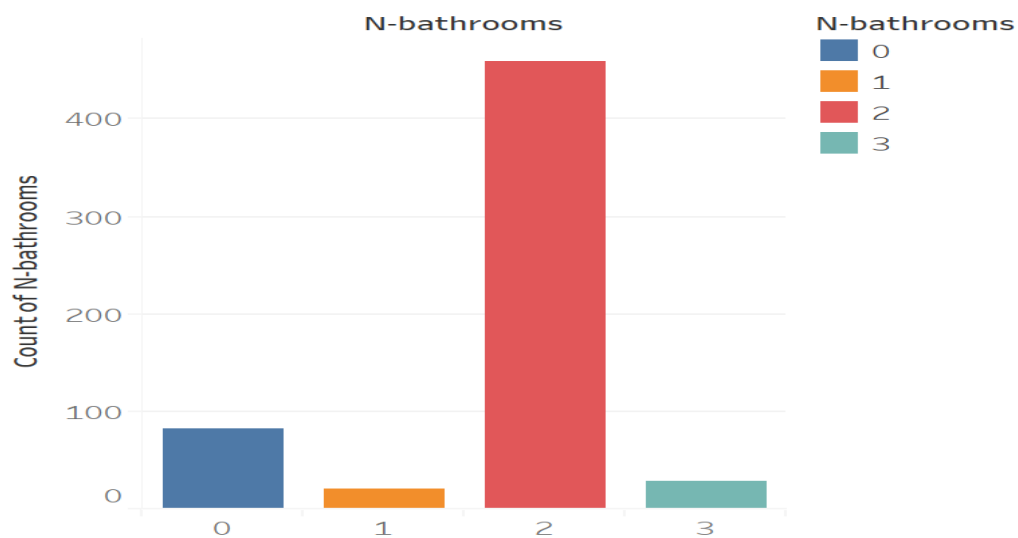


Figure 4.7 Distribution of data on categories from 0 to 3 in the N-bathrooms column.

N-kitchens

The following figure shows the distribution of data on categories from 0 to 4 in the kitchens column, where it found 82 records containing category 0, 497 records containing category 1 and 12 records containing category 2.

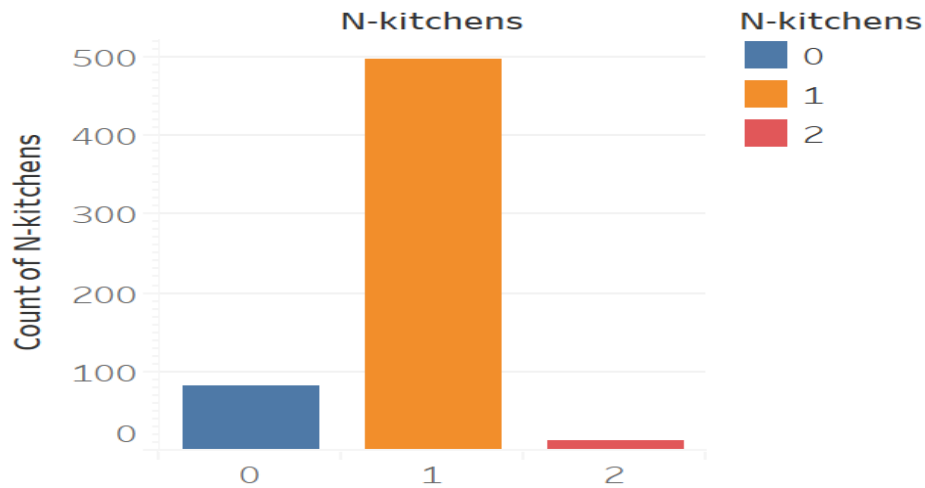


Figure 4.8 Distribution of data on categories from 0 to 4 in the kitchens column.

N-halls

The following figure shows the distribution of data on categories from 0 to 4 in the halls column, where it found 82 records containing category 0, 497 records containing category 1 and 12 records containing category 2.

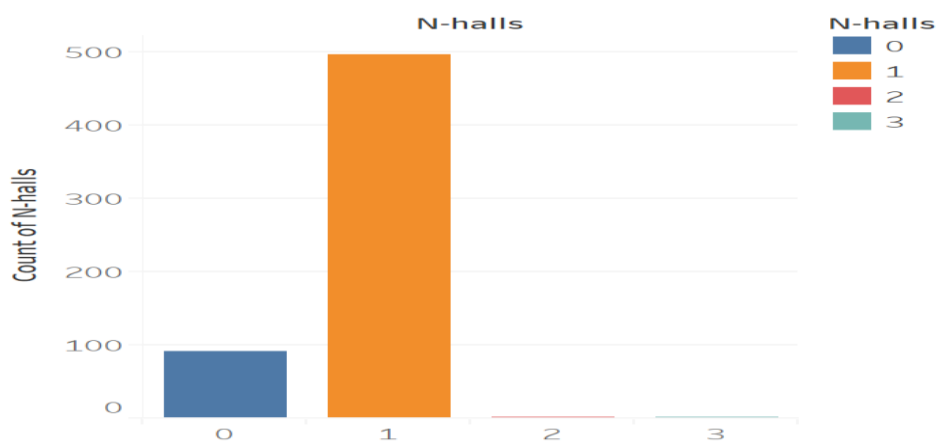


Figure 4.9 Distribution of data on categories from 0 to 4 in the halls column.

Sides

In this attribute, the side corresponding to the property is determined, and the values have been distributed to five categories, east, west, south, north, and open, the open represents the property that is open from all sides .The following figure shows the distribution of data on these categories, where there are 81 records on the northern side, 55 records in the southern side, and there are 147 record on the eastern side, and there are 152 records on the western side, and 156 records containing the value are open.

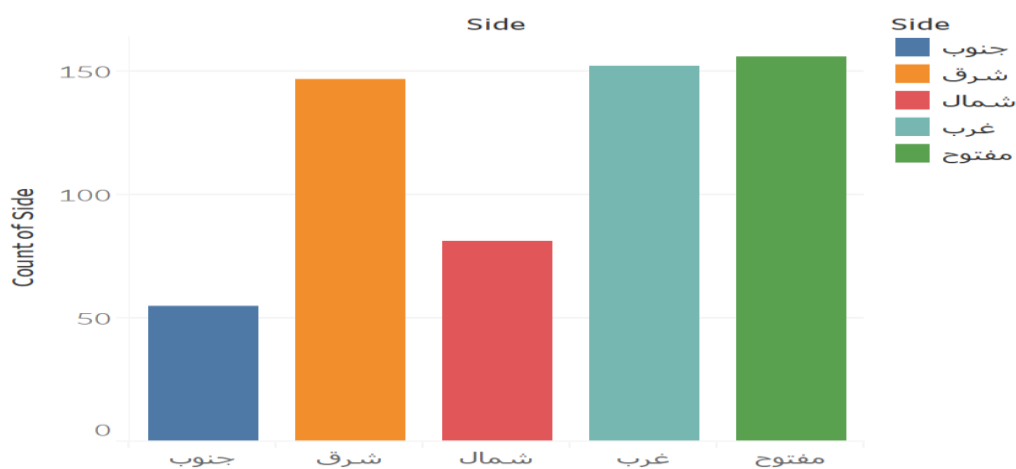


Figure 4.10 Distributed of data to five categories, east, west, south, north, and open.

Access-Road

In this attribute of the type of access road, the properties that the type of access road has a backfill, which numbered 364 records, and the number of properties that the type of access road were asphalt, numbered 227 records have noticed. The following figure shows the data by type of access road.

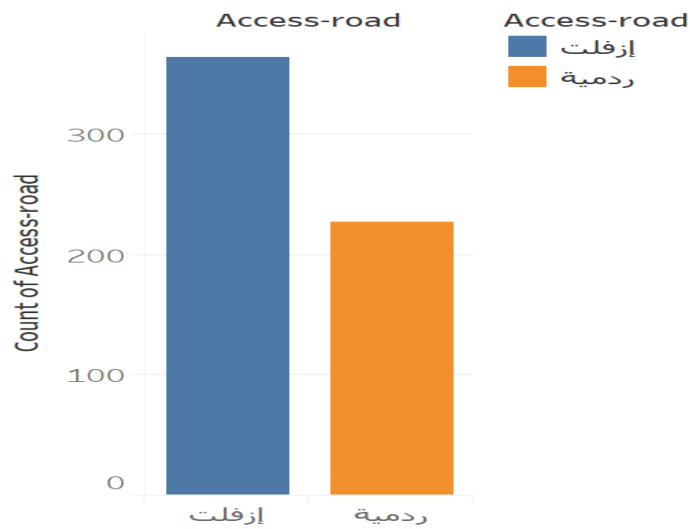


Figure 4.11 Type of access road.

Area

In this attribute, the real estate space is distributing into three categories small, medium, and large. The following figure shows the distribution of data on these categories, where there are 381 short category records, 187 medium category records, and 23 high category records.

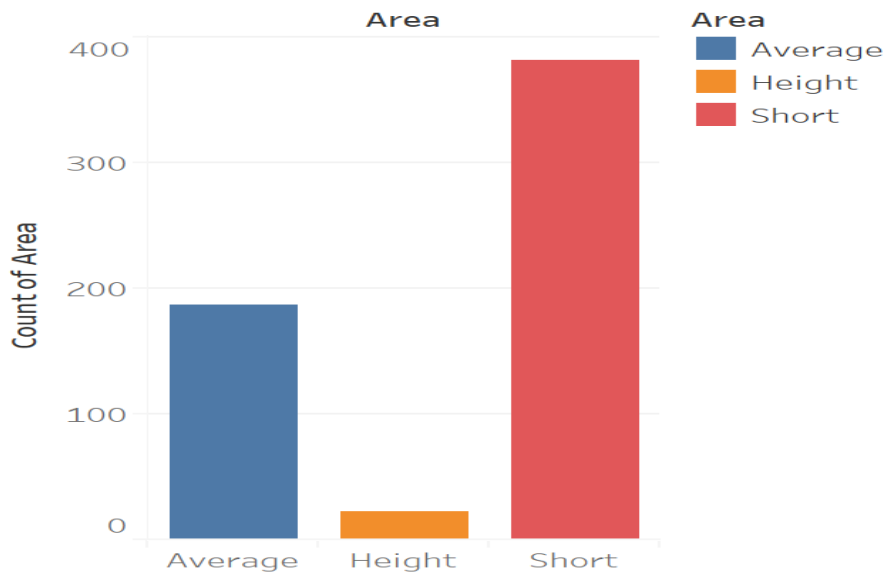


Figure 4.12 Space categories.

N-Rooms

The following figure shows the distribution of data on categories from 1 to 6 in the number of rooms column, where it found 66 records containing category 1 and expressing one room, 35 records containing category 2 and expressing two rooms, and 153 records containing category 3, which represent 3 rooms and 259 records contain category 4, which represents 4 rooms, 69 records contain category 5, which represents 5 rooms, and 9 records contain category 6, which represents 6 rooms.

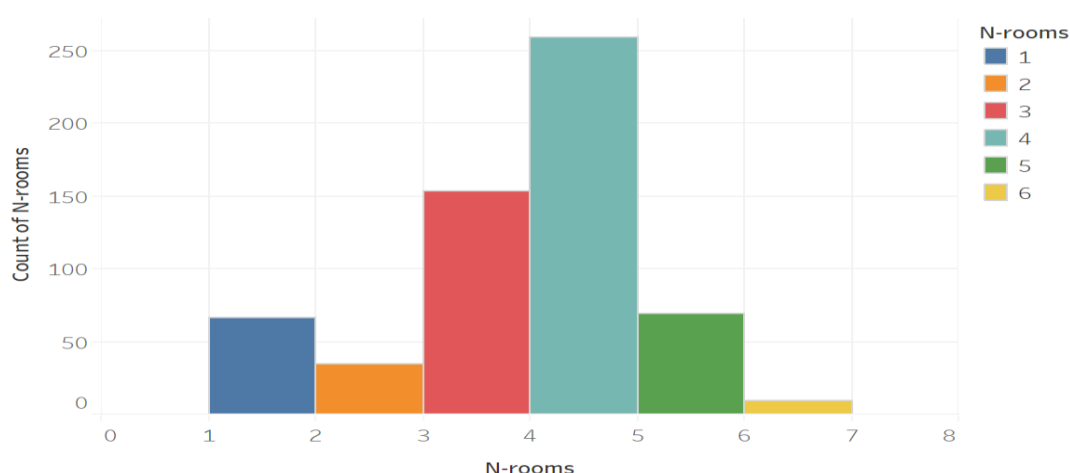


Figure 4.13 Distribution of data on categories from 1 to 6 in the number of rooms column.

Concrete/ non-concrete

In this figure, the number of properties whose construction type is concrete and the number of properties whose construction type is non-concrete, there are 575 concrete properties and 16 non-concrete properties.

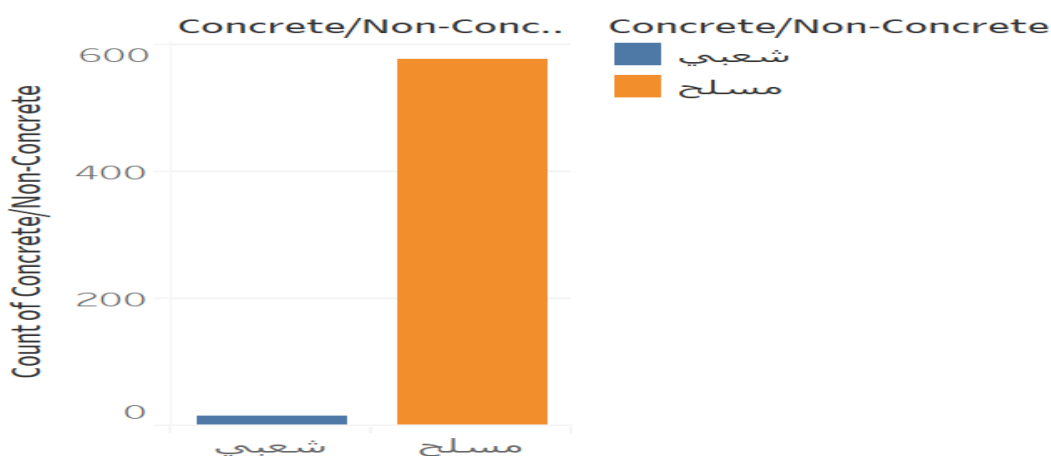


Figure 4.14 number of properties whose construction type is concrete and the number of properties whose construction type is non-concrete.

Families/ Singles

The data in this variable was distributed to three categories, which are “families” whether the property is available to families only, “singles” if the property is available to singles only, or “all” whether the property is available to all.

22 records available to singles only, 454 records available to families only, and 115 records available to all.

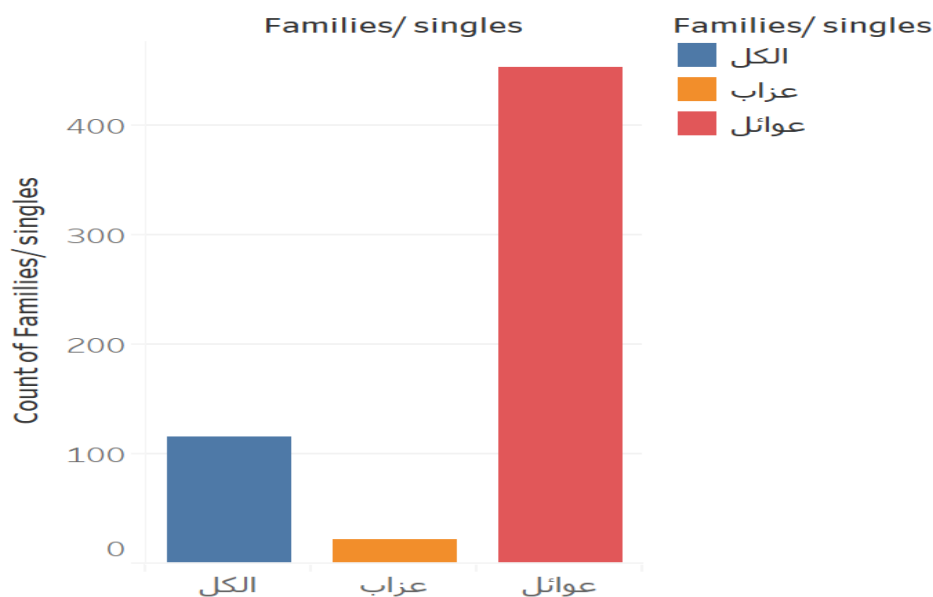


Figure 4.15 Families / Singles attribute.

Mountain/Marine

The data in this attribute is distributed in terms of the location of the property on the mountainous and marine categories, which shows if it is on the mountainous side of the main street or on the sea side of the main street. A number of 352 records were obtained on the mountainous side and 239 records on the sea side.

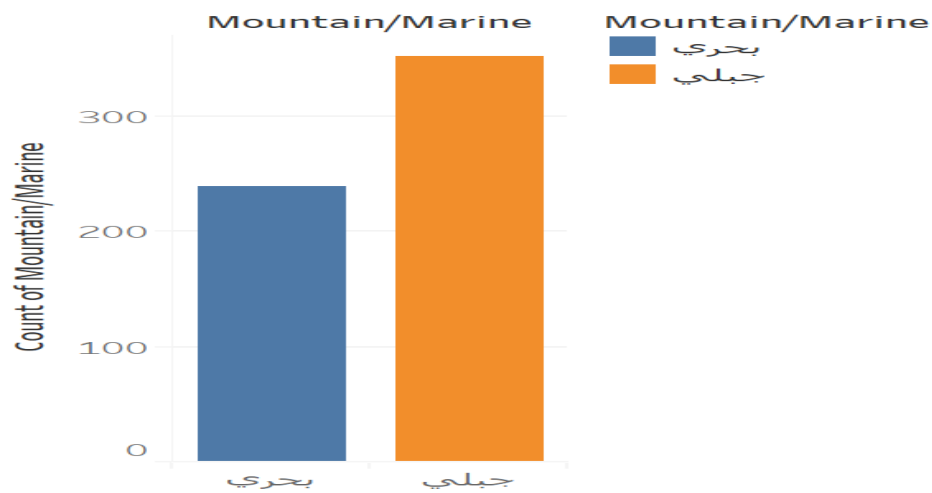


Figure 4.16 Mountainous and marine Categories.

Neighborhood

In the neighborhood attribute, the data was distributed to 4 areas, namely Ibn Sina, Al-Masaken, In Shaat and the Motdrreen. The number of records in the Ibn Sina area is 91 records, the number of properties in the Al-Masaken area is 152 records, the number of properties in the AL-Motdrreen area is 156 records, and the number of properties in the In-Shaat area is 192 records.

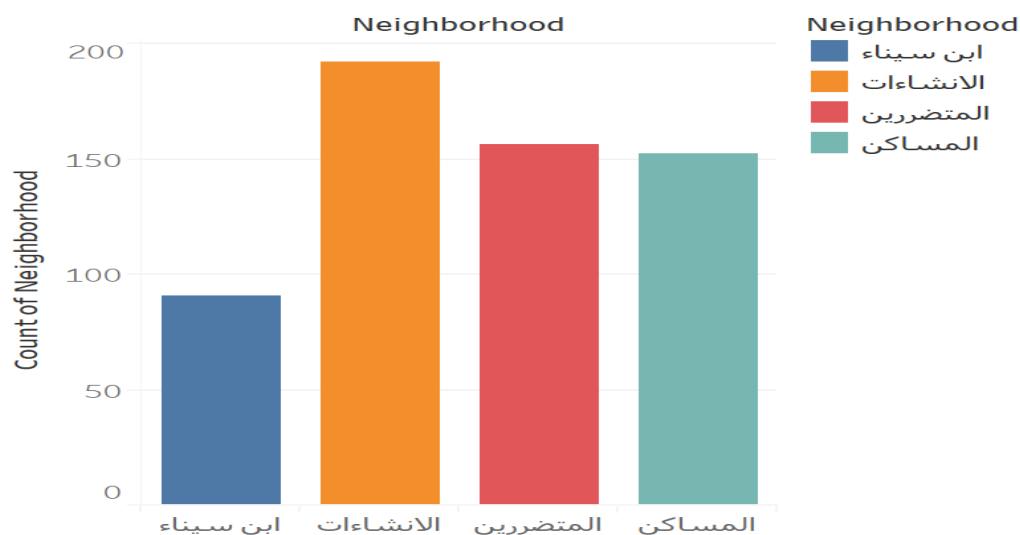


Figure 4.17 Neighborhood Categories.

Population-density

The data in this attribute is distributed to three categories A, B and C where A represents a high population density, there are 267 records that carry the value A and B represents a medium population density, and there are 219 records that carry the

value B and C represents a low population density and there are 109 records that carry the value C. It shows in the following figure:

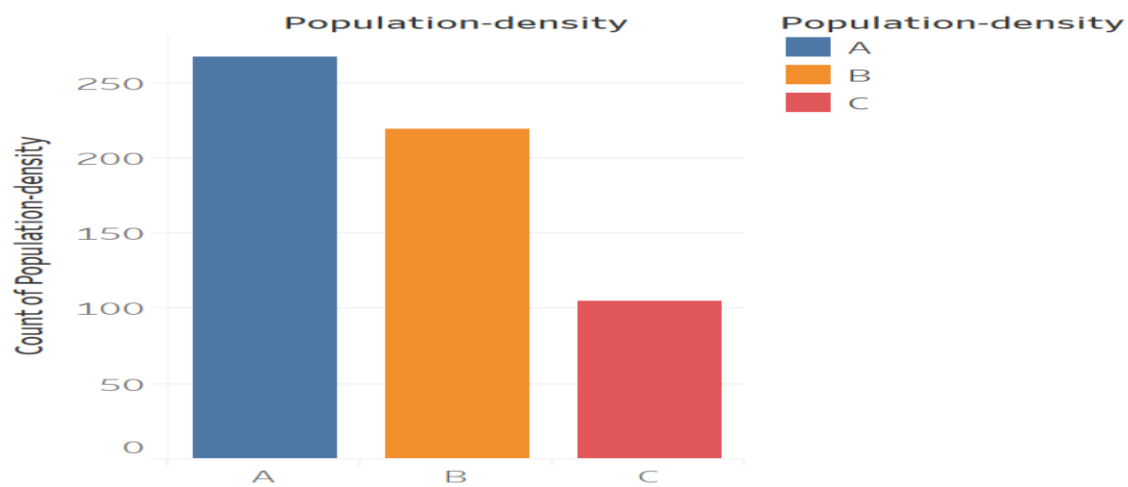


Figure 4.18 Population-density catrgorice.

Property-type

The following figure shows the distribution of data on the types of real estate collected, whether apartments or stores, 509 apartments and 82 stores were acquired.

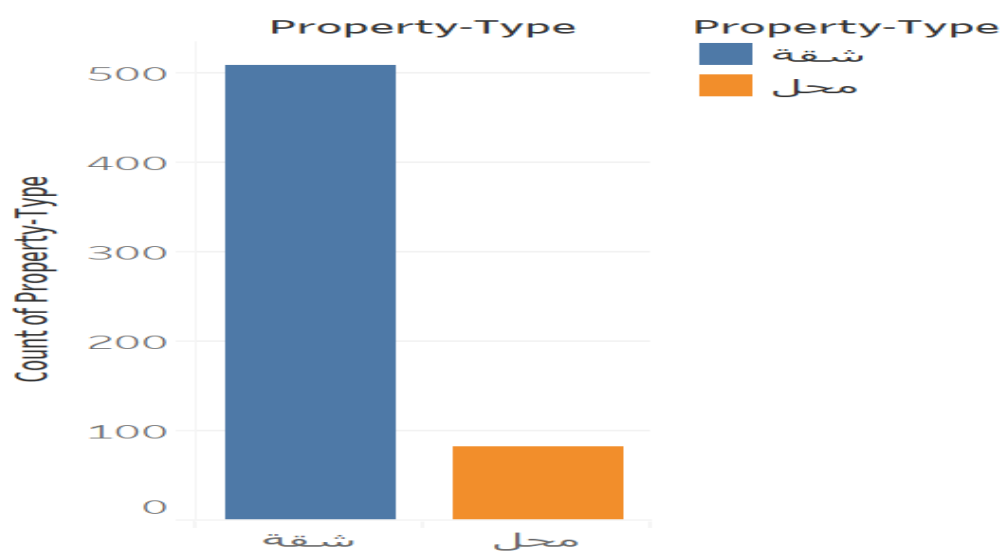


Figure 4.19: Property-type catrgorice.

Property-Age

In this figure, the number of properties whose real estate age is considered new and the number of properties whose age is considered old, and there are 349 new and 242 old.

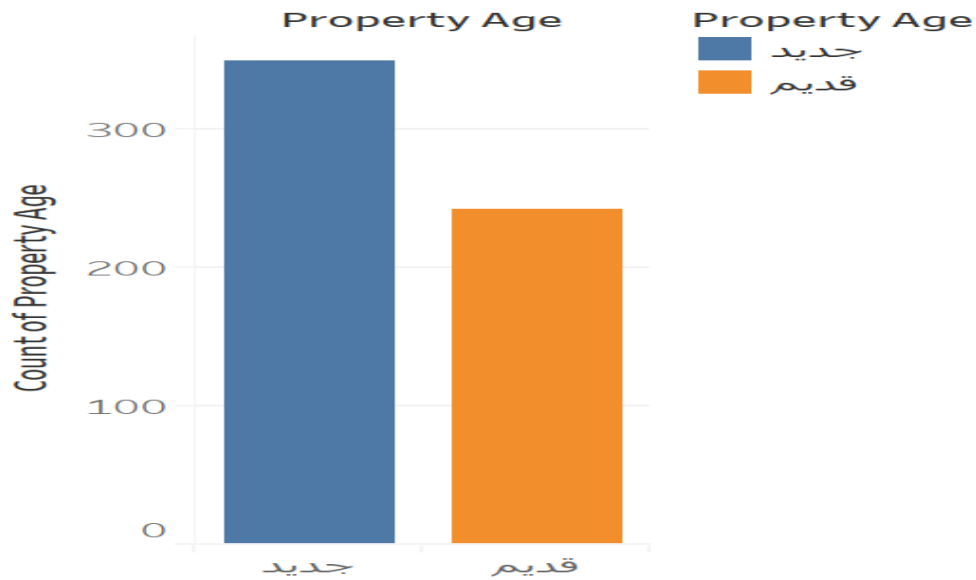


Figure 4.20 Property-Age categories.

Property Condition

In this figure, the number of rented properties and the number of non-rented properties are classified. There are 506 rented properties and 85 non-rented properties.

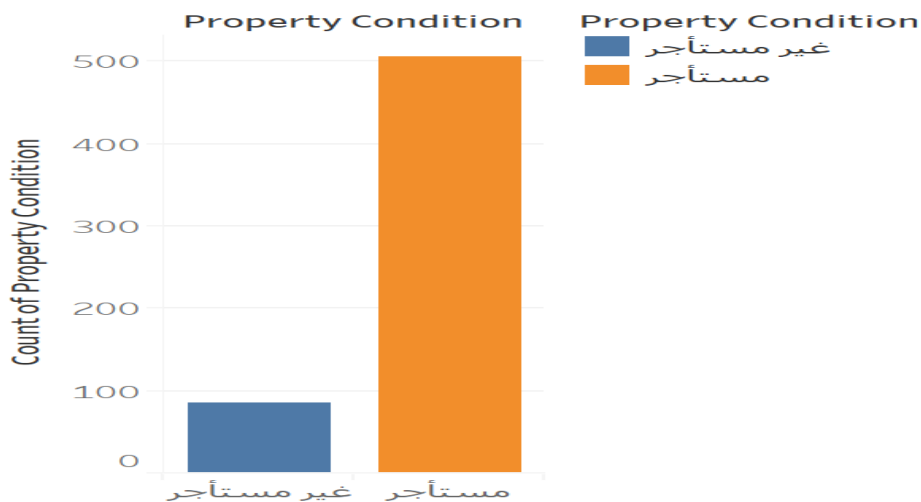


Figure 4.21 Property Condition categories.

Rental-period

In this figure, the number of the period of the old tenants, whether they were new tenants or old tenants, was clarified. As for the non-rented property, it was referred to as non-rented, and there are 355 new, 154 old and 28 non-rented.

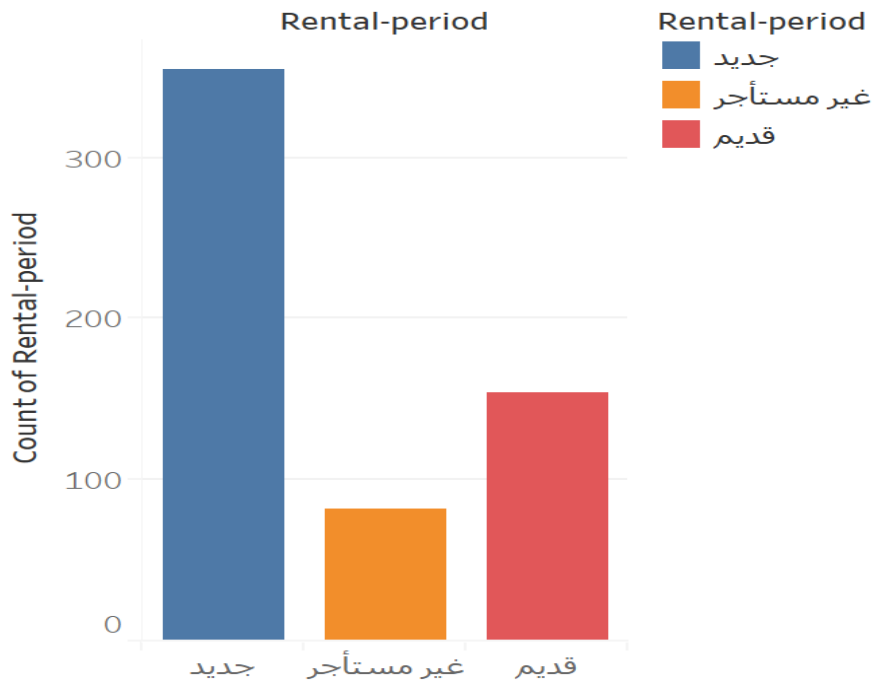


Figure 4.22 Rental-period categories.

Residential/Commercial

In this attribute, the data are distributing to three categories in terms of whether the property is used for a commercial or residential purpose or for both purposes. 464 records were obtained for a residential purpose only, 131 records are available for a commercial purpose only, and 14 are available for both cases.

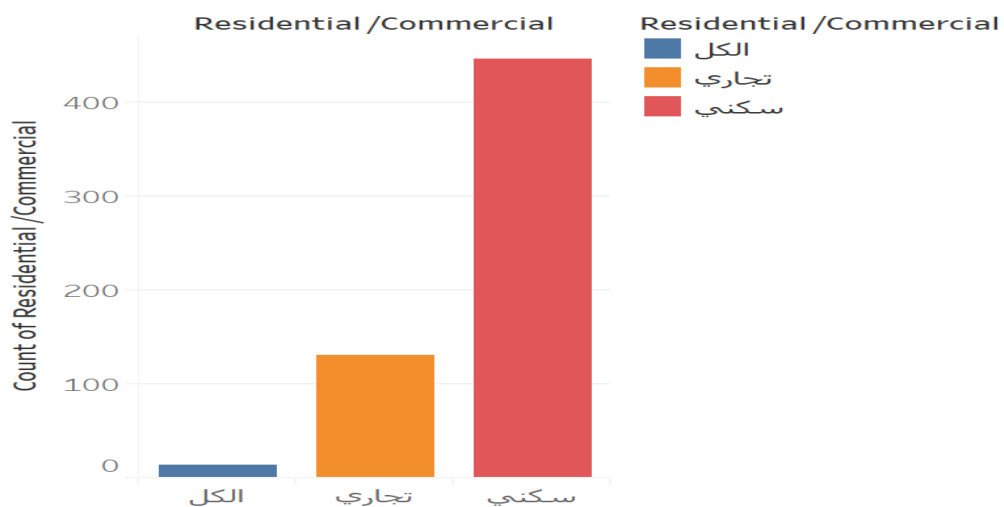


Figure 4.23 Residential/Commercial categories.

Services

In the services attribute, the data was divided into 3 categories: A, excellent services, B, medium services, and C, weak services. For category A, the number of records was 243, B, which had 348 records, and C, which had 0 records.



Figure 4.24 Services categories.

Street-type

In this figure, the data distribution has been clarified in terms of the street opposite the property, whether it is a main or sub-street, and there are 47 properties opposite the main street and 544 properties opposite of a sub-street.

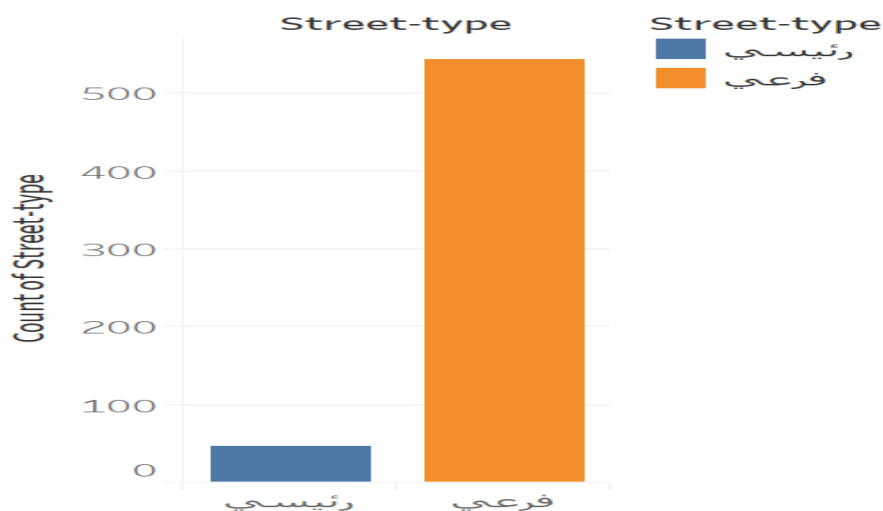


Figure 4.25 Street-type categories.

Water-meter

In this figure, the number of properties that have a private water meter and the number of properties in which the water meter is shared. It contains 64 shared water meters and 527 private water meters.

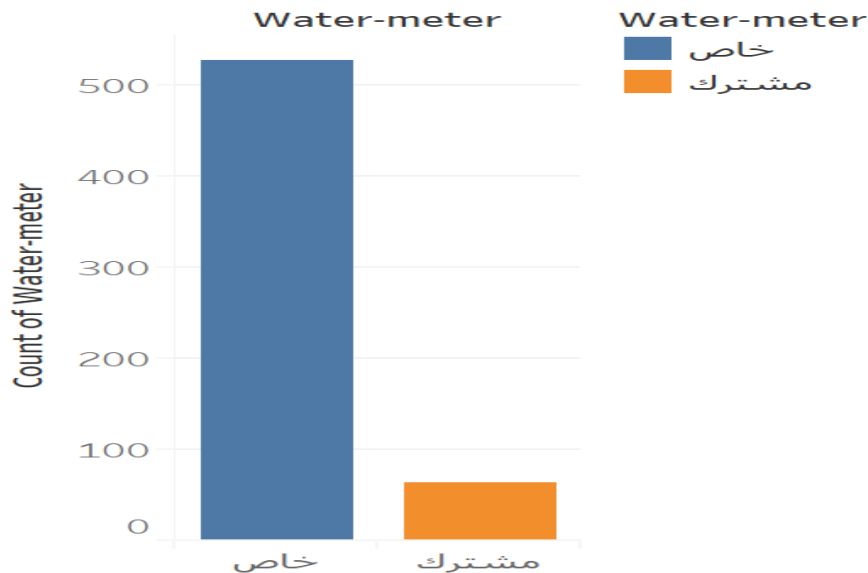


Figure 4.26 Water-meter categories.

Adjacent-Sides

The direction of each property has been determined, whether its direction is east, west, north or south, and if the property does not have adjacent properties, it includes all directions and was described as open. In category 0, there are 271 records, 208 in category 1, 48 records in category 2, and 64 records in category 3.

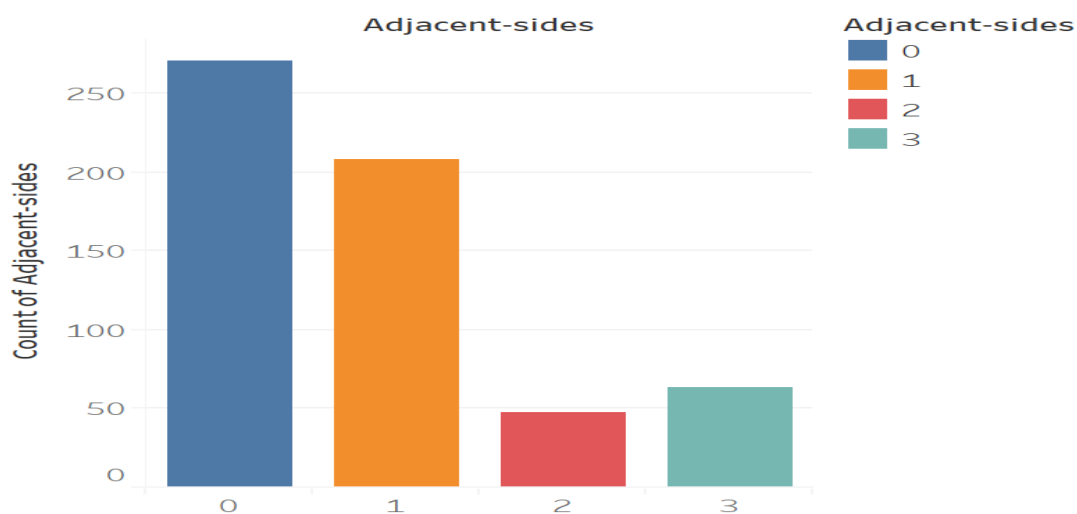


Figure 4.27 Adjacent-Sides categories.

Contract

In this figure, the number of properties for each type of contract was clarified, whether yearly, every two years, every three years, or every four years. The number of 567 records for yearly contract, 18 records for a two-year contract, 5 records for a three-year contract, and 1 record for a four-year contract.

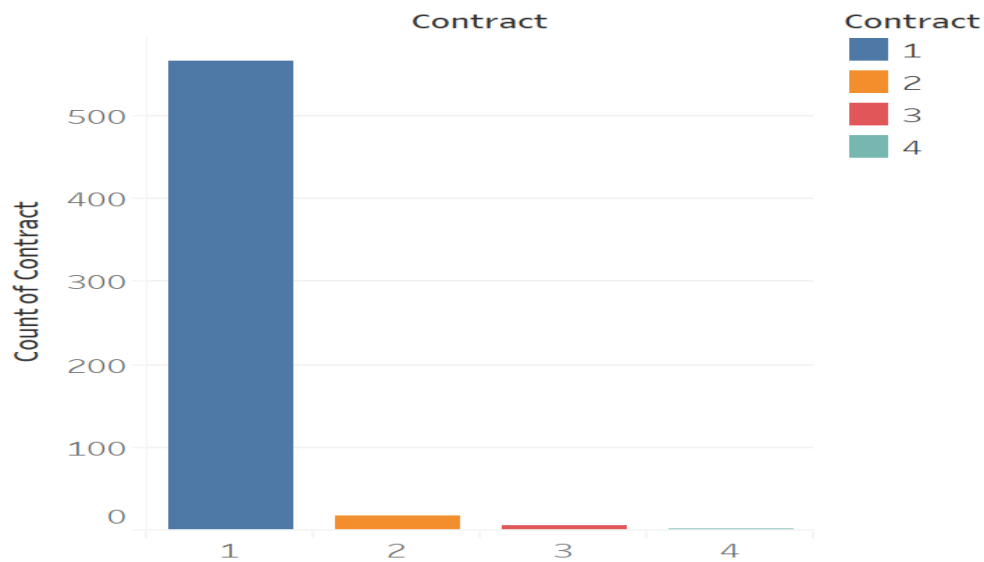


Figure 4.28 Contract categories.

N-entrances

The following figure shows the distribution of data on categories from 1 to 3 in the column of entry number, where it found 281 records containing category 1, 307 records containing category 2 and 3 records containing category 3.

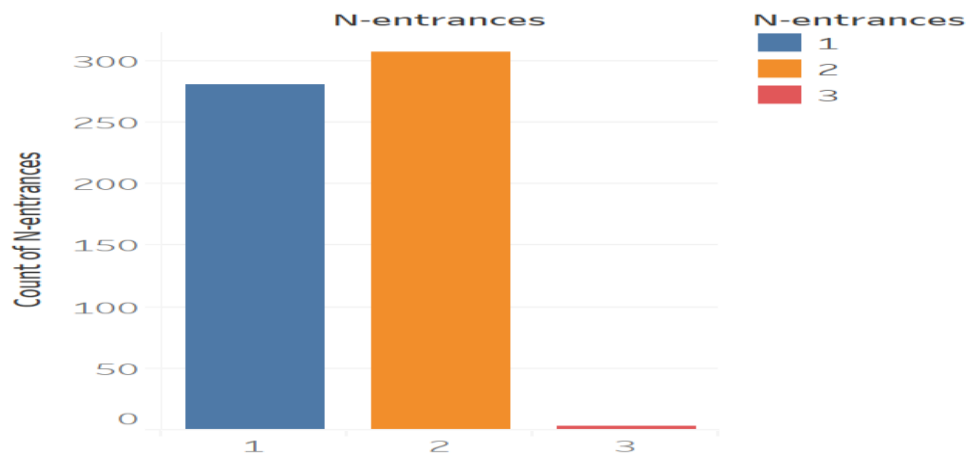


Figure 4.29 N-entrances categories.

N-floor

The following figure shows the distribution of data on categories from 1 to 6, where it found 145 records containing category 1, 174 records containing category 2, 161 records containing category 3, 83 records containing category 4, 22 records containing 5 and 6 records containing category 6.

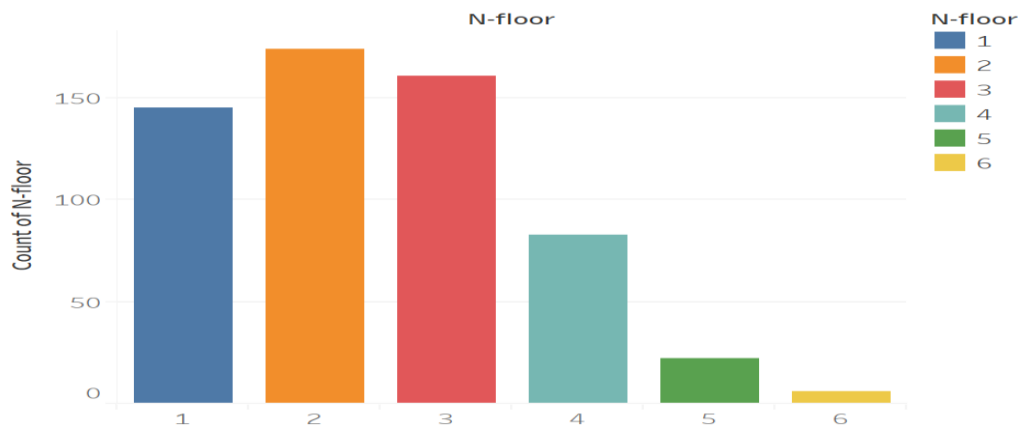


Figure 4.30 N-floor categories.

Price

The following figure shows the distribution of real estate prices in the data, where it was shown in the above figure that the average, which was shown in green color, is 602, and the large number of data is close to the average, which gives a good indication, however, there are some data that are relatively far from the average. As for the median value that was represented in blue color, which equals 573.0, which is somewhat close to the average of the data and its location, as for the most frequent value, which is called mode, which equals 500 as shown in the figure, as for the red color, it shows the value of the standard deviation, which is equal to 193 which is the value that represents the distance of the points from the mean, which is an average value, and this shows the possibility of obtaining a value close to the average when predicting the prices of the value of a particular property according to the values, which measure the dispersion of the data dispersion, and also the variance, which is one of the measures of dispersion, is equal to 37362.4 The STD is the root of the variance.

```

In [40]:
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
sns.displot(Data,x = "Price",bins=8,height=5), plt.xlabel("Price", size=14), plt.ylabel("Count", size=14)
plt.axvline(x=Data.Price.mean(),color='green'),plt.axvline(x=Data.Price.median(),color='blue')
plt.axvline(x=Data.Price.std(),color='red')
Out[40]:
<matplotlib.lines.Line2D at 0x23383231f40>

```

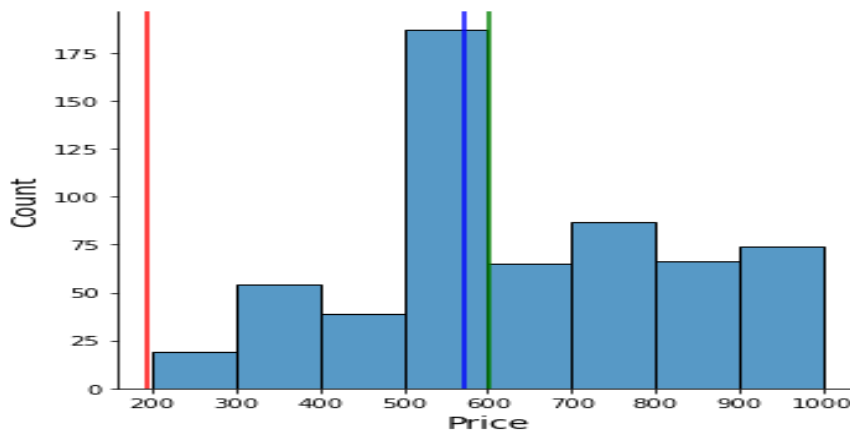


Figure 4.31:Price distribution histogram.

Figure (4.33) shows how the data is distributed and whether there are outliers, the yellow line represents the average value of the data and the bottom of this line is the first quartile (1st / 25th quartile) which is 500 which is the middle number between the smallest number (not the “minimum”) and the dataset mean, the part above the line is the third quartile (Q3/75th percentile) which is 700 which is the middle value between the median and the highest value (not the "maximum") of the data set and the interquartile range (IQR) is 25 to 75 percent, and the line which is called (whiskers) comes out from the bottom of the box to the smallest value which is 200 which represents $(Q1 - 1.5 * IQR)$ and the line that comes out from the top of the box to the largest value (whiskers) which is 1000 and represents $(Q3 + 1.5 * IQR)$, and $Q1 = 500$, about 25%, the percentage of the data is less than 500 and about 75%, the percentage is above 500.

```

In [44]:
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')

```

```
plt.boxplot(Data['Price'])
```

Out[44]:

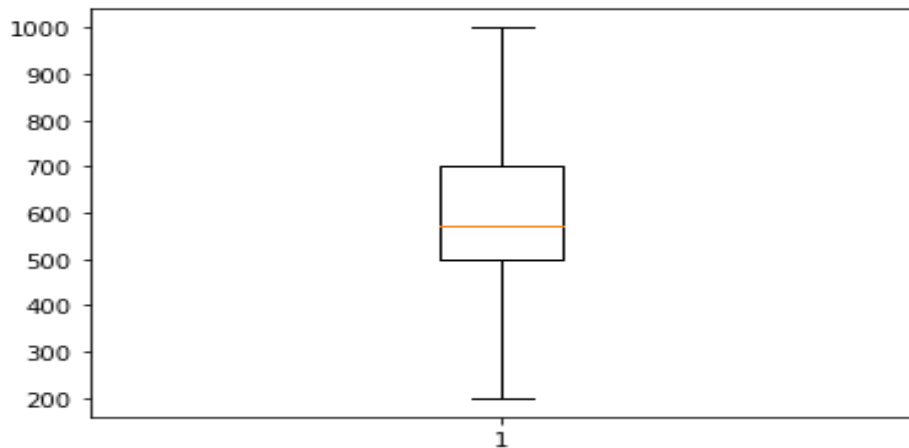


Figure 4.32:Price Boxplot

Distance

The following figure shows the distribution of the data with relative to the distance of the property from the main street. The average shown in green represents the average value, which is 314.1. Also, the mean is concentrated in an area where the amount of data is average, which is neither dense nor little , but there are areas such as the area from 0 to 100 that are considered dense, which is It is very far from the mean, so there is a relative dispersion because of the large difference between the data, and the median value represented in blue color, which is 300, which is very close to the mean, which is located in an area where the data density is average with other areas where the data is more concentrated, and the most frequently in the data is 350, which is called mode and is close to the mean and median. And after looking at the figure, from 0-100 there is a high amount of data, although mode, mean and median are in logic 300-350, so this data has a percentage of diaspora And the STD represented in red is considered close to the mean, meaning that the value of the STD is high, and this indicates the divergence of the data from the mean because the higher the value of the STD, the greater the data dispersion.

In [42]:

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
sns.displot(Data,x = "Distance",bins=10,height=5,plt.xlabel("Distance", size=14),plt.ylabel("Count", size=14)
plt.axvline(x=Data.Distance.mean(),color='green'),plt.axvline(x=Data.Distance.median(),color='blue')
plt.axvline(x=Data.Distance.std(),color='red')
```

Out[42]:

<matplotlib.lines.Line2D at 0x23383b17160>

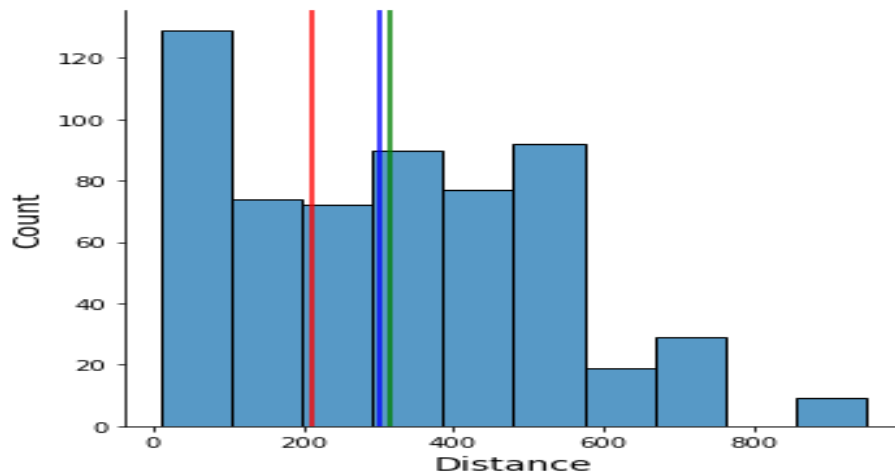


Figure 4.33:Distance distribution histogram.

In Figure (4.35), the first quartile (1st / 25th quartile) is 135, and the third quartile (Q3/75th percentile) is 480, the median value is 300, the lowest value is 10 and the largest value is 950.

In [41]:

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
plt.boxplot(Data['Distance'])
```

Out[41]:

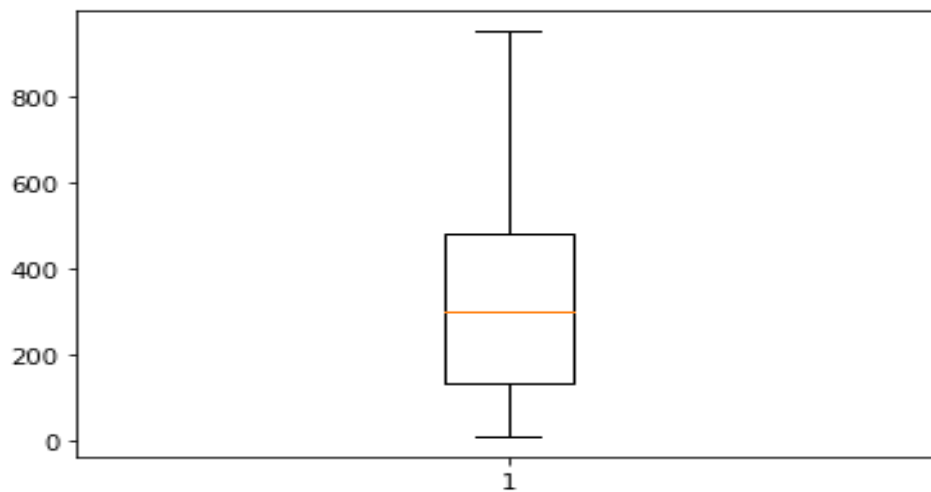


Figure 4.34: Distance Boxplot

4.4 Correlation

Correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. A value of ± 1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a – sign indicates a negative relationship. Usually, in statistics, four types of correlations are major.

4.4.1 Correlation matrix

To find the relationship between the continuous variables, the correlation matrix has used, in the collected data set there are 2 numerical variables, which are price and distance, and by executing the following code, the value of the relationship between the two variables is shown, which is a positive result of 0.66121. This value expresses that There is a weak relationship between the two variables.

In [1]:

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
In [3]:
Data[['Price','Distance']].corr()
```

Out[1]:

| | Price | Distance |
|----------|----------|----------|
| Price | 1.000000 | 0.066121 |
| Distance | 0.066121 | 1.000000 |

Figure 4.35: Price and Distance Correlation

To clarify the relationship between the two variables in the heatmap, run the following code:

```
In [6]:
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
In [7]:
dataplot=sns.heatmap(Data[['Price','Distance']].corr(), cmap="YlGnBu",annot=True)
plt.show()
```

The heatmap gives a clarification of the relationship between the variables, as well as to clarify the concentration of the data and the relationship between the variable price and distance in the following figure:

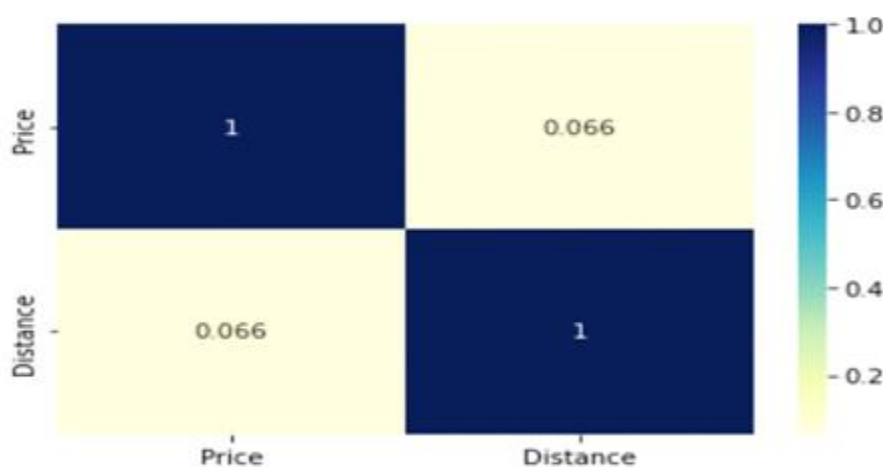


Figure 4.36 Correlation matrix heatmap

4.4.2 Analysis of Variance (ANOVA) for relationship between categorical and numerical variables

This test measures if there are any significant differences between the means of the values of the numeric variable for each categorical value.

Items must be remembered about ANOVA hypothesis test:

- Null hypothesis(H_0): The variables are not correlated with each other.
- P-value: The probability of null hypothesis being true.
- Accept Null hypothesis if $P\text{-value} > 0.05$. Means variables are not correlated.
- Reject Null hypothesis if $P\text{-value} < 0.05$. Means variables are correlated.

As the output of the P-value is almost zero, hence, the H_0 has rejected. Which means the variables are correlated with each other.

To find the relationship between a numerical variable and a categorical variable, ANOVA test has used and in the case of testing the numerical variable price with the categorical variable property- condition, used the following code:

```
In [5]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [6]:
CategoryGroupLists=Data.groupby('Property Condition')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.04094476452310667

The result shows that the two variables have a strong relationship, which is illustrated by the value of the variable p-value which is less than 0.05.

To test the relationship between the numerical variable price and the categorical variable deluxe/standard, the following code has used:

```
In [3]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
```

```

from scipy.stats import f_oneway
In [4]:
CategoryGroupLists=Data.groupby('Deluxe/ Standard')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])

```

P-Value for Anova is: 1.001801851167343e-10

The result shows that the relationship is somewhat weak between the two variables because the result of the variable p-value is greater than 0.05.

In order to test the relationship between the numerical variable price and the categorical variable adjacent-sides, the following code was used:

```

In [7]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [8]:
CategoryGroupLists=Data.groupby('Adjacent-sides')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
P-Value for Anova is: 0.048761515052840466

```

And the result shows that the relationship is strong between the two variables because the result of variable p-value is less than 0.05.

As for the relationship between the numerical variable price with the categorical variable street -type, it turned out that the relationship is very strong between the two variables, because the result is 0.02, which is shown below.

```

In [9]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [10]:
CategoryGroupLists=Data.groupby('Street-type')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])

```

P-Value for Anova is: 0.023288873920696337

As for the relationship between the numerical variable price with the categorical variable services, it turns out that it is not strong, but at the same time it is not weak,

and through the following code, the result 0.08 has gotten, which is considered somewhat acceptable.

```
In [11]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [12]:
CategoryGroupLists=Data.groupby('Services')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.08762257609987782

As for the relationship between the numerical variable price and the categorical variable population-density, it turns out that the relationship is strong, and through the following code that was implemented, the result 0.02 has gotten, which is very excellent.

```
In [13]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [14]:
CategoryGroupLists=Data.groupby('Population-density')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.02289035821369028

As for the relationship between the numerical variable price with the categorical variable, the type of access-road, it turns out that it is not strong, but at the same time it is not weak, and through the implemented code, the result 0.08 has gotten, which is considered somewhat acceptable.

```
In [13]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [15]:
CategoryGroupLists=Data.groupby('Access-road')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.08426815579686779

To test the relationship between the numerical variable price and the categorical variable of the water-meter, the following code has used:

```
In [13]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [16]:
CategoryGroupLists=Data.groupby('Water-meter')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 1.8981009506056453e-07

And the result shows that the relationship is weak between the two variables because the result of the variable p-value is greater than 0.05.

As for the relationship between the numerical variable price and the categorical variable concrete/Nonconcrete, it turns out that it is weak and very strong between the two variables, the result 2.23 has gotten, which is shown below.

```
In [13]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [17]:
CategoryGroupLists=Data.groupby('Concrete/Non-Concrete')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
P-Value for Anova is: 2.2332050905519324e-05
```

Regarding the relationship between the numerical variable price with the categorical variable singles/families, it turned out that the relationship is weak between the two variables, the result 0.16 has gotten, which is shown below.

```
In [13]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [18]:
CategoryGroupLists=Data.groupby('Families/ singles')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.16497031093466583

Regarding the relationship between the numerical variable price with the categorical variable residential-commercial, it turned out that the relationship is weak between the two variables, the result 0.99 has gotten, which is shown below.

```
In [13]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [19]:
CategoryGroupLists=Data.groupby('Residential/Commercial')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.5991290062251479

As for the relationship between the numerical variable price and the categorical variable side, it turned out that the relationship is weak between the two variables.

```
In [13]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [21]:
CategoryGroupLists=Data.groupby('Side')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 1.064643780447361e-05

With regard to the relationship between the numerical variable price with the categorical variable contract, it turned out that the relationship is weak between the two variables.

```
In [13]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [22]:
CategoryGroupLists=Data.groupby('Contract')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.40545091249034426

To test the relationship between the numerical variable price and the categorical variable rental -period, the following code has used:

In [13]:

```
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
```

In [23]:

```
CategoryGroupLists=Data.groupby('Rental-period')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.0075584054971972675

And the result shows that the relationship is strong between the two variables because the result of variable p-value is less than 0.05.

To test the relationship between the numerical variable price and the categorical variable mountain/marine, the following code has used:

In [13]:

```
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
```

In [24]:

```
CategoryGroupLists=Data.groupby('Mountain/Marine')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.0004348063134586595

The result shows that the relationship is strong between the two variables, and the result is 0.00. This value is ideal and strong, and it is the best value has obtained in the data.

With regard to the relationship between the numerical variable price with the categorical variable property- type, it turned out that the relationship is weak between the two variables, and through the result from implementing the code, which is 0.85, the weakness of this relationship was clarified.

In [13]:

```
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
```

```

from scipy.stats import f_oneway
In [25]:
CategoryGroupLists=Data.groupby('Property-Type')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])

```

P-Value for Anova is: 0.8572024136850738

With regard to the relationship between the numerical variable price with the categorical variable number of rooms, it turned out that the relationship is weak between the two variables, the result 1.1 has gotten, which is shown below.

```

In [13]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [26]:
CategoryGroupLists=Data.groupby('N-rooms')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])

```

P-Value for Anova is: 1.1041071867516625e-11

As for the relationship between the numerical variable price with the categorical variable number of bathrooms, it turned out that the relationship is weak between the two variables, the result 3.04 has gotten, which is shown below.

```

In [13]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [27]:
CategoryGroupLists=Data.groupby('N-bathrooms')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])

```

P-Value for Anova is: 3.0400352515871166e-12

As for the relationship between the numerical variable price with the categorical variable number of kitchens, it turned out that the relationship is weak between the two variables, the result 6.04 has gotten, which is shown below.

```

In [13]:
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway

```

In [28]:

```
CategoryGroupLists=Data.groupby('N-kitchens')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 2.9980484461374486e-06

As for the relationship between the numerical variable price with the categorical variable number of entries, it turned out that the relationship is weak between the two variables, the result 6.04 has gotten, which is explained below.

In [13]:

```
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
```

In [29]:

```
CategoryGroupLists=Data.groupby('N-entrances')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 6.04816608467149e-24

As for the relationship between the numerical variable price with the categorical variable number of halls, it turned out that the relationship is weak between the two variables ,which is explain below :

In [13]:

```
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
```

In [30]:

```
CategoryGroupLists=Data.groupby('N-halls')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.19833707464819977

As for the relationship between the numerical variable price and the categorical variable the number of floors, it became clear that the relationship is weak between the two variables, the result 0.126 has gotten, which is shown below.

In [13]:

```
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
```

In [31]:


```
CategoryGroupLists=Data.groupby('N-floor')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.12618430771421596

To test the relationship between the numerical variable price with the categorical variable number of balconies, the following code has used:

In [13]:

```
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [33]:
CategoryGroupLists=Data.groupby('N-balconies')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 0.027104167140616926

The result shows that the relationship is strong between the two variables, and the result is 0.02. This value is considered ideal and strong.

With regard to the relationship between the numerical variable price and the categorical variable area, it turned out that the relationship is weak between the two variables, and result is 1.155, which is shown below.

In [13]:

```
import pandas as pd
Data=pd.read_csv('RealEstate_Dataset.CSV')
from scipy.stats import f_oneway
In [34]:
CategoryGroupLists=Data.groupby('Area')['Price'].apply(list)
AnovaResults=f_oneway(*CategoryGroupLists)
print('P-Value for Anova is:',AnovaResults[1])
```

P-Value for Anova is: 1.1557797969017152e-05

Chapter Five

Experimental and Discussion Modeling

5.1 Introduction

In this chapter the common technologies and the evaluation matrix used for model evaluation are presented , Later the result and discussing is described.

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring [32].

Accuracy Measurements: Accuracy simply means the number of values correctly predicted. The model accuracy can measured by two methods.

5.1.1 Confusion Matrix

A confusion matrix is a tabular summary of the number of correct and incorrect predictions made by a classifier. It is used to measure the performance of a classification model. It can be used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1-score [33].

Confusion matrices are widely used because they give a better idea of a model's performance than classification accuracy does. For example, in classification accuracy, there is no information about the number of misclassified instances. Imagine that the data has two classes where 85% of the data belongs to class A, and 15% belongs to class B. Also, assume that the classification model correctly classifies all the instances of class A, and misclassifies all the instances of class B. In this case, the model is 85% accurate. However, class B is misclassified, which is undesirable. The confusion matrix, on the other hand, displays the correctly and incorrectly classified instances for all the classes and will, therefore, give a better insight into the performance of the classifier [33].

The following four are the basic terminology which will help in determining the needed metrics.

- **True Positives (TP):** when the actual value is Positive and predicted is also Positive.
- **True negatives (TN):** when the actual value is Negative and prediction is also Negative.
- **False positives (FP):** When the actual value is negative but prediction is Positive. Also known as the Type 1 error.
- **False negatives (FN):** When the actual value is Positive but the prediction is Negative. Also known as the Type 2 error.

For a binary classification problem, there is a 2 x 2 matrix as shown below with four values:

| | | Actual Values | |
|------------------|--------------|---------------|--------------|
| | | Positive (1) | Negative (0) |
| Predicted Values | Positive (1) | TP | FP |
| | Negative (0) | FN | TN |

Figure 5.1: 2x2 Confusion Matrix [33]

- The target variable has two values: Positive or Negative.
- The columns represent the actual values of the target variable.
- The rows represent the predicted values of the target variable.

5.1.2 Classification Measure

Basically, it is an extended version of the confusion matrix. There are measures other than the confusion matrix which can help achieve better understanding and analysis of the model and its performance.

- a. TP Rate.
- b. FP Rate.
- c. Accuracy.
- d. Precision.
- e. Recall (equivalent to TP Rate).
- f. F1-Score.

a. TP Rate: rate of true positives (instances correctly classified as a given class).

$$TPR = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN}$$

Equation 5.1 TP Rate [33]

b. FP Rate: rate of false positives (instances falsely classified as a given class).

$$FPR = \frac{FP}{Actual\ Negative} = \frac{FP}{TN + FP}$$

Equation 5.2 FP Rate [33]

c. Accuracy: Accuracy simply measures how often the classifier makes the correct prediction. It's the ratio between the number of correct predictions and the total number of predictions. It is a measure of correctness that is achieved in true

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{Correct\ Predictions}{Total\ Predictions}$$

Equation 5.3 Accuracy [33]

prediction. In simple words, it tells us how many predictions are actually positive out of all the total positive predicted.

d. Precision: Precision is defined as the ratio of the total number of correctly classified positive classes divided by the total number of predicted positive classes. Or, out of all the predictive positive classes, how much values have predicted correctly. Precision should be high(ideally 1).

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{Predictions Actually Positive}}{\text{Total Predicted positive}}$$

Equation 5.4 Precision [33]

e. Recall: It is a measure of actual observations which are predicted correctly, i.e. how many observations of positive class are actually predicted as positive. It is also known as Sensitivity. *Recall* is a valid choice of evaluation metric which is captures *as many positives* as possible.

Recall is defined as the ratio of the total number of *correctly classified positive classes* divide by the *total number of positive classes*. Or, out of all the positive classes, how much values have predicted correctly. Recall should be high(ideally 1).

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{Predictions Actually Positive}}{\text{Total Actual positive}}$$

Equation 5.5 Recall [33]

f. F-measure / F1-Score: The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall. The harmonic mean is used because it is not sensitive to extremely large values, unlike simple averages.

F1 score sort of maintains a balance between the precision and recall for the classifier. If the precision is low, the F1 is low and if the recall is low again the F1 score is low.

$$\text{F1-Score} = 2 * \frac{(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

Equation 5.6 F-measure / F1-Score [33]

5.2 Experiment and result

Here some of the most common used algorithms are comprised and evaluated by using weak software.

WEKA-an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools to develop machine learning techniques and apply them to real-world data mining problems [34].

5.2.1 Experiment Models

1-JRip Algorithm

=== Stratified cross-validation ===

=== Summary ===

| | | |
|----------------------------------|-----|-----------|
| Correctly Classified Instances | 507 | 85.7868 % |
| Incorrectly Classified Instances | 84 | 14.2132 % |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|--------------|
| | 0.943 | 0.227 | 0.829 | 0.943 | 0.882 | 0.733 | 0.873 | 0.818 | '(334-650]' |
| | 0.833 | 0.049 | 0.913 | 0.833 | 0.871 | 0.798 | 0.910 | 0.910 | '(650-inf)' |
| | 0.391 | 0.007 | 0.818 | 0.391 | 0.529 | 0.543 | 0.772 | 0.398 | '(-inf-334]' |
| Weighted Avg. | 0.858 | 0.142 | 0.860 | 0.858 | 0.851 | 0.743 | 0.879 | 0.821 | |

=== Confusion Matrix ===

```

a b c <-- classified as
300 15 3 | a = '(334-650]'
```

```

37 189 1 | b = '(650-inf)'
```

```

25 3 18 | c = '(-inf-334]'
```

2-MulityClassClassifier Algorithm

=== Stratified cross-validation ===

=== Summary ===

| | | |
|----------------------------------|-----|-----------|
| Correctly Classified Instances | 436 | 73.7733 % |
| Incorrectly Classified Instances | 155 | 26.2267 % |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|--------------|
| | 0.802 | 0.300 | 0.757 | 0.802 | 0.779 | 0.505 | 0.845 | 0.851 | '(334-650]' |
| | 0.753 | 0.184 | 0.718 | 0.753 | 0.735 | 0.565 | 0.875 | 0.842 | '(650-inf)' |
| | 0.217 | 0.011 | 0.625 | 0.217 | 0.323 | 0.341 | 0.892 | 0.527 | '(-inf-334]' |
| Weighted Avg. | 0.738 | 0.233 | 0.732 | 0.738 | 0.727 | 0.515 | 0.860 | 0.822 | |

=== Confusion Matrix ===

```

a b c <-- classified as
255 61 2 | a = '(334-650]'
```



```
52 171 4 | b = '(650-inf)'
30 6 10 | c = '(-inf-334]'
```

3-LWL Algorithm

=== Stratified cross-validation ===

=== Summary ===

| | | |
|----------------------------------|-----|-----------|
| Correctly Classified Instances | 426 | 72.0812 % |
| Incorrectly Classified Instances | 165 | 27.9188 % |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|--------------|
| | 0.965 | 0.549 | 0.672 | 0.965 | 0.792 | 0.495 | 0.832 | 0.850 | '(334-650]' |
| | 0.493 | 0.036 | 0.896 | 0.493 | 0.636 | 0.545 | 0.855 | 0.816 | '(650-inf)' |
| | 0.152 | 0.004 | 0.778 | 0.152 | 0.255 | 0.325 | 0.822 | 0.352 | '(-inf-334]' |
| Weighted Avg. | 0.721 | 0.310 | 0.766 | 0.721 | 0.691 | 0.501 | 0.840 | 0.798 | |

=== Confusion Matrix ===

a b c <-- classified as

```
307 9 2 | a = '(334-650]'
```

```
115 112 0 | b = '(650-inf)'
```

```
35 4 7 | c = '(-inf-334]'
```

4-MultilayerPerceptron Algorithm

=== Stratified cross-validation ===

=== Summary ===

| | | |
|----------------------------------|-----|-----------|
| Correctly Classified Instances | 512 | 86.6328 % |
| Incorrectly Classified Instances | 79 | 13.3672 % |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|--------------|
| | 0.918 | 0.179 | 0.856 | 0.918 | 0.886 | 0.745 | 0.956 | 0.963 | '(334-650]' |
| | 0.894 | 0.066 | 0.894 | 0.894 | 0.894 | 0.828 | 0.974 | 0.964 | '(650-inf)' |
| | 0.370 | 0.011 | 0.739 | 0.370 | 0.493 | 0.497 | 0.879 | 0.607 | '(-inf-334]' |
| Weighted Avg. | 0.866 | 0.123 | 0.862 | 0.866 | 0.859 | 0.758 | 0.957 | 0.936 | |

=== Confusion Matrix ===

```

a  b  c  <-- classified as
292 21  5 |  a = '(334-650]'
```

```

23 203  1 |  b = '(650-inf)'
```

```

26  3 17 |  c = '(-inf-334]'
```

5-NaiveBayes Algorithm

=== Stratified cross-validation ===

=== Summary ===

| | | |
|----------------------------------|-----|-----------|
| Correctly Classified Instances | 374 | 63.2826 % |
| Incorrectly Classified Instances | 217 | 36.7174 % |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|--------------|
| | 0.670 | 0.304 | 0.720 | 0.670 | 0.694 | 0.365 | 0.759 | 0.795 | '(334-650]' |
| | 0.652 | 0.173 | 0.701 | 0.652 | 0.676 | 0.486 | 0.761 | 0.732 | '(650-inf)' |
| | 0.283 | 0.130 | 0.155 | 0.283 | 0.200 | 0.117 | 0.781 | 0.175 | '(-inf-334]' |
| Weighted Avg. | 0.633 | 0.240 | 0.669 | 0.633 | 0.648 | 0.392 | 0.762 | 0.722 | |

=== Confusion Matrix ===

```

a  b  c  <-- classified as
213 57 48 |  a = '(334-650]'
```

```

56 148 23 |  b = '(650-inf)'
```

```
27 6 13 | c = '(-inf-334]'
```

6-BayestNet Algorithm

=== Stratified cross-validation ===

=== Summary ===

| | | |
|----------------------------------|-----|-----------|
| Correctly Classified Instances | 378 | 63.9594 % |
| Incorrectly Classified Instances | 213 | 36.0406 % |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|--------------|
| | 0.814 | 0.535 | 0.640 | 0.814 | 0.716 | 0.300 | 0.694 | 0.683 | '(334-650]' |
| | 0.480 | 0.170 | 0.637 | 0.480 | 0.548 | 0.332 | 0.723 | 0.688 | '(650-inf)' |
| | 0.217 | 0.009 | 0.667 | 0.217 | 0.328 | 0.355 | 0.652 | 0.272 | '(-inf-334]' |
| Weighted Avg. | 0.640 | 0.354 | 0.641 | 0.640 | 0.621 | 0.317 | 0.702 | 0.653 | |

=== Confusion Matrix ===

```
a b c <-- classified as
```

```
259 57 2 | a = '(334-650]'
```

```
115 109 3 | b = '(650-inf)'
```

```
31 5 10 | c = '(-inf-334]'
```

7-REP Tree Algorithm

=== Stratified cross-validation ===

=== Summary ===

| | | |
|----------------------------------|-----|---------|
| Correctly Classified Instances | 503 | 85.11 % |
| Incorrectly Classified Instances | 88 | 14.89 % |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|--|---------|---------|-----------|--------|-----------|-----|----------|----------|-------|
|--|---------|---------|-----------|--------|-----------|-----|----------|----------|-------|

| | | | | | | | | |
|---------------|-------|-------|-------|-------|-------|-------|-------|--------------|
| 0.940 | 0.227 | 0.828 | 0.940 | 0.881 | 0.729 | 0.925 | 0.929 | '(334-650]' |
| 0.837 | 0.063 | 0.892 | 0.837 | 0.864 | 0.784 | 0.950 | 0.934 | '(650-inf)' |
| 0.304 | 0.006 | 0.824 | 0.304 | 0.444 | 0.479 | 0.860 | 0.454 | '(-inf-334]' |
| Weighted Avg. | 0.851 | 0.147 | 0.852 | 0.851 | 0.840 | 0.731 | 0.930 | 0.894 |

=== Confusion Matrix ===

a b c <-- classified as

299 17 2 | a = '(334-650]'

36 190 1 | b = '(650-inf)'

26 6 14 | c = '(-inf-334]'

5.2.2 Result

| | Criteria | JRIP | | MultyClass Classifier | | REP Tree | | LWL | | Multilayer Perceptron | | NaiveBayes | | BayestNet | |
|-----|----------------------------------|-------|--------|-----------------------|--------|----------|--------|-------|--------|-----------------------|--------|------------|--------|-----------|--------|
| Avg | Correctly Classified Instances | 507 | 85.78% | 436 | 73.77% | 503 | 85.11% | 426 | 72.08% | 512 | 86.63% | 374 | 63.28% | 374 | 63.95% |
| | Incorrectly Classified Instances | 84 | 14.21% | 155 | 26.22% | 88 | 14.89% | 165 | 27.91% | 79 | 13.36% | 217 | 36.71% | 213 | 36.04% |
| | TP Rate | 0.858 | | 0.738 | | 0.851 | | 0.721 | | 0.866 | | 0.633 | | 0.640 | |
| | FP Rate | 0.142 | | 0.233 | | 0.147 | | 0.310 | | 0.123 | | 0.240 | | 0.354 | |
| | Precision | 0.860 | | 0.732 | | 0.852 | | 0.766 | | 0.862 | | 0.669 | | 0.641 | |
| | Recall | 0.858 | | 0.738 | | 0.851 | | 0.721 | | 0.866 | | 0.633 | | 0.640 | |
| | F-Measure | 0.851 | | 0.727 | | 0.840 | | 0.691 | | 0.859 | | 0.648 | | 0.621 | |

Table 5.1:summarizes the result of the comparison and evaluation process

The table above summarizes the result of the comparison and evaluation process using different measurements , the first row represent the number and the percentage of correctly classified data, where MLP algorithm recorded the height percentage value with 86.63% followed by JRIP algorithm with 85.87%

The second row represents the number and the percentage of the incorrectly classified data ,which the highest value represent the worst algorithm, where the

NaïveBayes percentage is 36.04% which is the highest and worst value in all algorithms.

The third row represents the average of the True Positive Rate, where the MLP algorithm got the highest value with 0.866, followed by the JRIP algorithm with 0.858, and the last is the NaïveBayes algorithm, which scored the lowest rate with 0.640.

As for the fourth column, which is the false positive rate, in which the highest value is the worst, the NaïveBayes algorithm recorded the highest and worst value with 0.354.

The three next rows represent the Precision, Recall and F-Measure where highest values represent better performance and the MLP has the highest values with 0.862 for the Precision, 0.866 for the Recall and 0.859 for the F-Measure. It is followed by the JRIP algorithm with a very close ratio, with the following values 0.86 for the Precision, 0.858 for the Recall and 0.851 for the F-Measure, followed by the REP Tree algorithm with the following values 0.852 for the Precision, 0.851 for the Recall and 0.84 for the F-Measure.

Figure (5.2) shows the percentage of data that was correctly classified as well as the data that was incorrectly classified in each algorithm, where the green color represents the data that was correctly classified in each algorithm and the red color represents the data that was incorrectly classified in each algorithm.

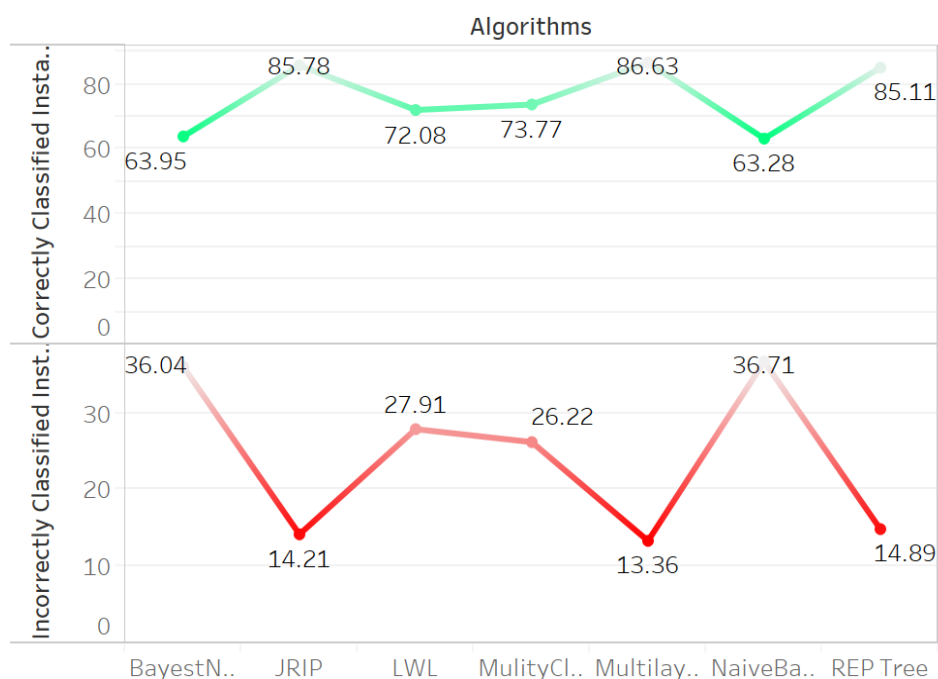


Figure 5. 2 Data incorrectly/correctly classified in MLP algorithm

This figure (5.3) represent the average recall of TP and FP rate for each algorithm, where green represents the TP rate and red represents the FP Rate.

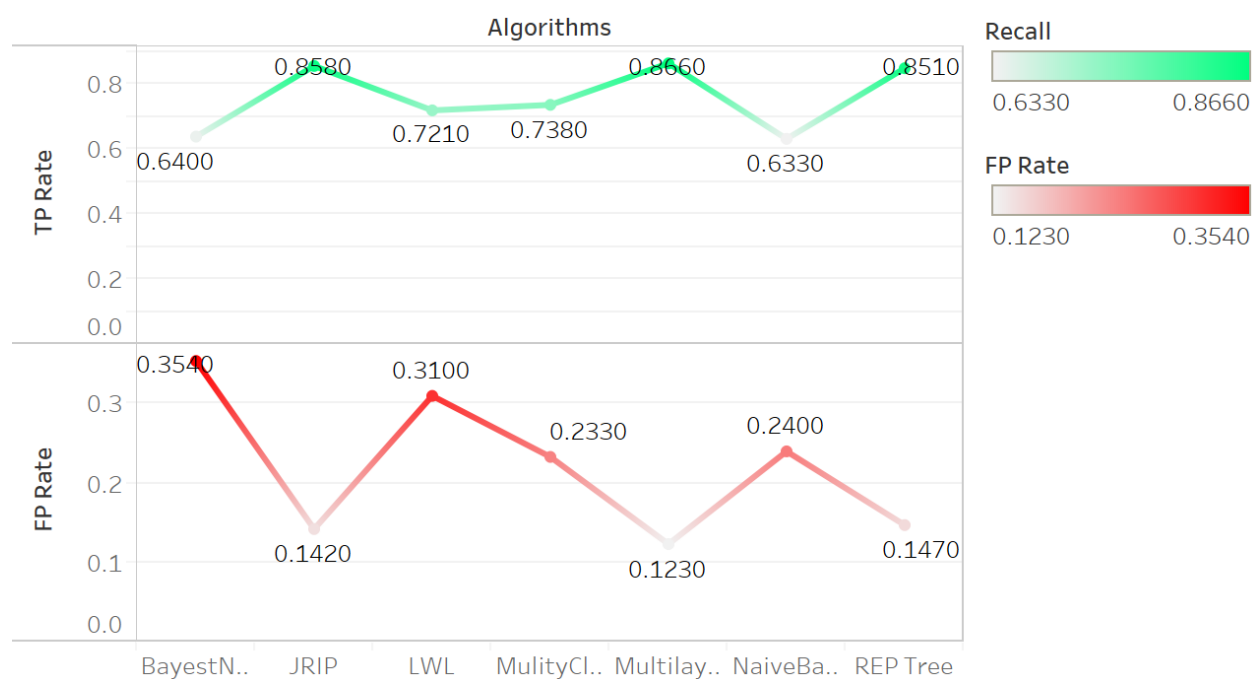


Figure 5. 3 the average recall rate and FP rate for each algorithm

Figure (5.4) shows the average of the F-Measuer and Precision and recall for each algorithm where the green color represents the average of the F-Measuer and the orange color represents the average of the Precision and the blue color represents the average of the Recall.

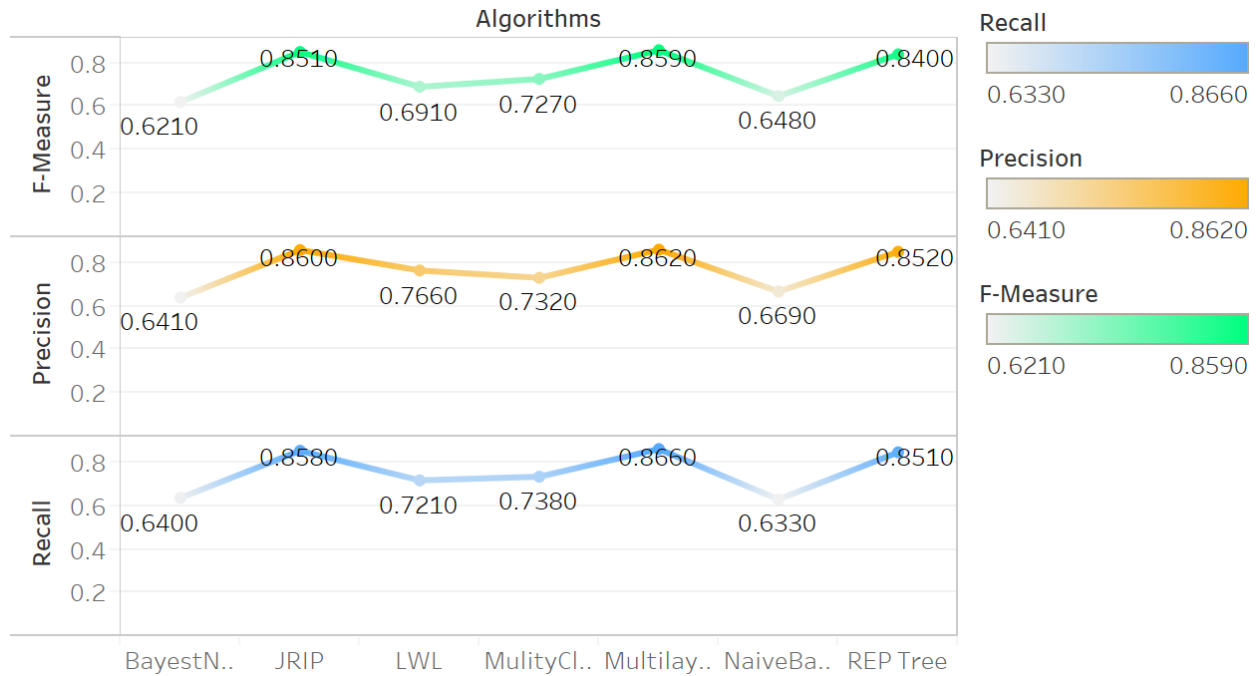


Figure 5. 4 F-Measuer, Precision and Recall for each algorithm

Chapter Six

Model Implementation

6.1 introduction

This chapter presents an overview of the Multilayer Perceptron (MLP) algorithm and how it works, and the implementation and visualization of the algorithm.

MLP are feedforward artificial neural networks that generate outputs from a set of inputs. In a Multilayer Perceptron, multiple layers of input nodes are connected as a directed graph between the input and output layers. The Multilayer Perceptron is a deep learning method that uses backpropagation to train the network [35].

MLPs are widely recognized as algorithms, they were originally designed for image recognition. It gets its name from performing the human-like function of perceiving, seeing, and identifying images [35].

MLPs are essentially feed-forward neural networks with three types of layers: input, output, and hidden. The input layer receives the input signal for processing. The output layer performs tasks such as classification and prediction. MLPs are accurate computational engine consists of an arbitrary number of hidden layers between input and output layers. Similarly, the data flow from the input layer to the output layer in a Multilayer Perceptron. The neurons in the MLPs are trained using the backpropagation learning algorithm. MLPs are designed to approximate any continuous function and can solve problems that are not linearly separable [35].

The Perceptron consists of an input layer and an output layer which are fully connected. MLPs have the same input and output layers but may have multiple hidden layers in between the aforementioned layers, as seen below [36].

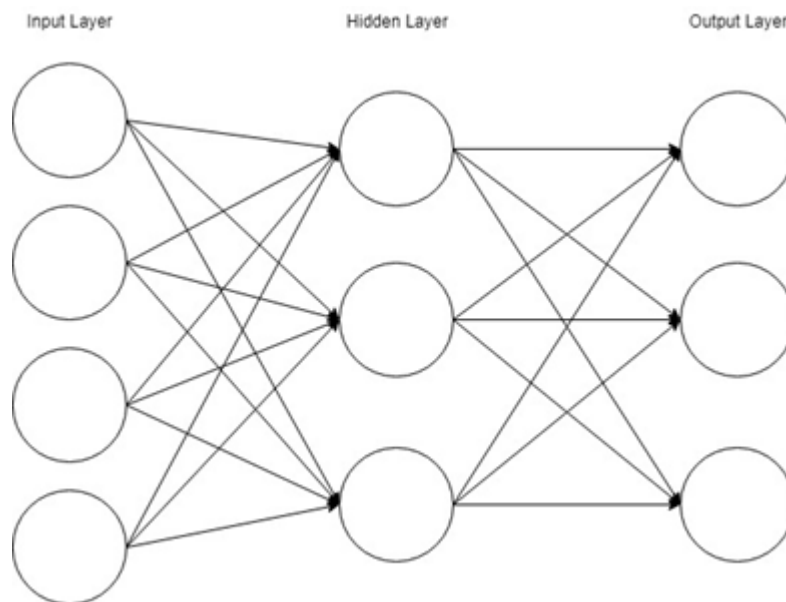


Figure 6. 1 layers perceptron

How does MLP algorithm work:

Step 1: Just as with the perceptron, the inputs are pushed forward through the MLP by taking the dot product of the input with the weights that exist between the input layer and the hidden layer ($W \rightarrow H$). This dot product yields a value at the hidden layer. This value does not pushed forward as with a perceptron though [36].

Step 2: MLPs utilize activation functions at each of their calculated layers. There are many activation functions to discuss: rectified linear units (ReLU), sigmoid function, tanh. Push the calculated output at the current layer through any of these activation functions [36].

Step 3: Once the calculated output at the hidden layer has been pushed through the activation function, push it to the next layer in the MLP by taking the dot product with the corresponding weights [36].

Step 4: Repeat steps two and three until the output layer is reached.

At the output layer, the calculations will either be used for a backpropagation algorithm that corresponds to the activation function that was selected for the MLP (in the case of training) or a decision will be made based on the output (in the case of testing) [36].

MLPs form the basis for all neural networks and have greatly improved the power of computers when applied to classification and regression problems. Computers are no longer limited by XOR cases and can learn rich and complex models because of the multilayer perceptron [36].

6.2 MLP implementation

Implementation of the multilayer perceptron algorithm in this project :

Step 1: Import the necessary libraries.

```
In [9]:
import numpy as np
import pandas as pd
from sklearn.neural_network import MLPClassifier
import ipywidgets as widg
```

Step 2: Upload the dataset.

```
In [10]:
Data = pd.read_csv("RealEstateDataset.csv")
```

Step 3: Convert categorical values into numerical values.

```
In [11]:
Data.PropertyType = pd.Categorical(Data.PropertyType)
Data['PropertyTypeENCODE'] = Data.PropertyType.cat.codes
Data.Country = pd.Categorical(Data.Country)
Data['CountryENCODE'] = Data.Country.cat.codes
Data.Governorate = pd.Categorical(Data.Governorate)
Data['GovernorateENCODE'] = Data.Governorate.cat.codes
Data.City = pd.Categorical(Data.City)
Data['CityENCODE'] = Data.City.cat.codes
Data.Neighborhood = pd.Categorical(Data.Neighborhood)
Data['NeighborhoodENCODE'] = Data.Neighborhood.cat.codes
Data.MountainMarine = pd.Categorical(Data.MountainMarine)
Data['MountainMarineENCODE'] = Data.MountainMarine.cat.codes
Data.NRooms = pd.Categorical(Data.NRooms)
Data['NRoomsENCODE'] = Data.NRooms.cat.codes
Data.NBathrooms = pd.Categorical(Data.NBathrooms)
Data['NBathroomsENCODE'] = Data.NBathrooms.cat.codes
Data.NKitchens = pd.Categorical(Data.NKitchens)
Data['NKitchensENCODE'] = Data.NKitchens.cat.codes
Data.NHalls = pd.Categorical(Data.NHalls)
Data['NHallsENCODE'] = Data.NHalls.cat.codes
Data.NFloor = pd.Categorical(Data.NFloor)
Data['NFloorENCODE'] = Data.NFloor.cat.codes
Data.NBalconies = pd.Categorical(Data.NBalconies)
Data['NBalconiesENCODE'] = Data.NBalconies.cat.codes
Data.PropertyAge = pd.Categorical(Data.PropertyAge)
Data['PropertyAgeENCODE'] = Data.PropertyAge.cat.codes
Data.RentalPeriod = pd.Categorical(Data.RentalPeriod)
Data['RentalPeriodENCODE'] = Data.RentalPeriod.cat.codes
Data.Contract = pd.Categorical(Data.Contract)
```

```

Data['ContractENCODE']=Data.Contract.cat.codes
Data.PropertyCondition=pd.Categorical(Data.PropertyCondition)
Data['PropertyConditionENCODE']=Data.PropertyCondition.cat.codes
Data.Side=pd.Categorical(Data.Side)
Data['SideENCODE']=Data.Side.cat.codes
Data.ResidentialCommercial=pd.Categorical(Data.ResidentialCommercial)
Data['ResidentialCommercialENCODE']=Data.ResidentialCommercial.cat.codes
Data.AccessRoad=pd.Categorical(Data.AccessRoad)
Data['AccessRoadENCODE']=Data.AccessRoad.cat.codes
Data.PopulationDensity=pd.Categorical(Data.PopulationDensity)
Data['PopulationDensityENCODE']=Data.PopulationDensity.cat.codes
Data.Services=pd.Categorical(Data.Services)
Data['ServicesENCODE']=Data.Services.cat.codes
Data.Distance=pd.Categorical(Data.Distance)
Data['DistanceENCODE']=Data.Distance.cat.codes
Data.StreetType=pd.Categorical(Data.StreetType)
Data['StreetTypeENCODE']=Data.StreetType.cat.codes
Data.AdjacentSides=pd.Categorical(Data.AdjacentSides)
Data['AdjacentSidesENCODE']=Data.AdjacentSides.cat.codes
Data.DeluxeStandard=pd.Categorical(Data.DeluxeStandard)
Data['DeluxeStandardENCODE']=Data.DeluxeStandard.cat.codes
Data.FurnitureNonFurnished=pd.Categorical(Data.FurnitureNonFurnished)
Data['FurnitureNonFurnishedENCODE']=Data.FurnitureNonFurnished.cat.codes
Data.Area=pd.Categorical(Data.Area)
Data['AreaENCODE']=Data.Area.cat.codes

```

Step 4: Deleted the old attribute.

```

In [12]:
EncodedData=Data.drop(['PropertyType','Country','Governorate',
                        'City','Neighborhood','MountainMarine',
                        'NRooms','NBathrooms','NKitchens','NHalls',
                        'NFloor','NBalconies','PropertyAge',
                        'RentalPeriod','Contract','PropertyCondition',
                        'Side','ResidentialCommercial','FamiliesSingles',
                        'ConcreteNonConcrete','NEntrances','WaterMeter',
                        'AccessRoad','PopulationDensity','Services',

                        'Distance','StreetType','AdjacentSides','DeluxeStandard',
                        'FurnitureNonFurnished','Area'], axis=1)

```

Step 5: Create model and fit the data into the created model.

```

In [13]:
X = EncodedData.drop(columns="Price")
# Create model object
model = MLPClassifier()
# Fit data onto the model
model.fit(X,Y)

```

Step 6: Create user input fields.

```

In [14]:
PropertyType=widg.FloatText()
Country=widg.FloatText()
Governorate=widg.FloatText()
City=widg.FloatText()

```

```

Neighborhood=widg.FloatText()
MountainMarine=widg.FloatText()
NRooms=widg.FloatText()
NBathrooms=widg.FloatText()
NKitchens=widg.FloatText()
NHalls=widg.FloatText()
NFloor=widg.FloatText()
NBalconies=widg.FloatText()
PropertyAge=widg.FloatText()
RentalPeriod=widg.FloatText()
Contract=widg.FloatText()
PropertyCondition=widg.FloatText()
Side=widg.FloatText()
ResidentialCommercial=widg.FloatText()
AccessRoad=widg.FloatText()
PopulationDensity=widg.FloatText()
Services=widg.FloatText()
Distance=widg.FloatText()
StreetType=widg.FloatText()
Adjacentsides=widg.FloatText()
DeluxeStandard=widg.FloatText()
FurnitureNonFurnished=widg.FloatText()
Area=widg.FloatText()
PropertyType.description="PropertyType"
Country.description="Country"
Governorate.description="Governorate"
City.description="City"
Neighborhood.description="Neighborhood"
MountainMarine.description="MountainMarine"
NRooms.description="NRooms"
NBathrooms.description="NBathrooms"
NKitchens.description="NKitchens"
NHalls.description="NHalls"
NFloor.description="NFloor"
NBalconies.description="NBalconies"
PropertyAge.description="PropertyAge"
RentalPeriod.description="RentalPeriod"
Contract.description="Contract"
PropertyCondition.description="PropertyCondition"
Side.description="Side"
ResidentialCommercial.description="ResidentialCommercial"
AccessRoad.description="AccessRoad"
PopulationDensity.description="PopulationDensity"
Services.description="Services"
Distance.description="Distance"
StreetType.description="StreetType"
Adjacentsides.description="Adjacentsides"
DeluxeStandard.description="DeluxeStandard"
FurnitureNonFurnished.description="FurnitureNonFurnished"
Area.description="Area"

display(PropertyType,Country,Governorate,City,Neighborhood,MountainMarine,
        NRooms,NBathrooms,NKitchens,NHalls,NFloor,NBalconies,PropertyAge,
        RentalPeriod,Contract,PropertyCondition,Side,ResidentialCommercial,
        AccessRoad,PopulationDensity,Services,Distance,StreetType,Adjacentsides,
        DeluxeStandard,FurnitureNonFurnished,Area)

```

| | | | |
|---------------|--------------------------------|----------------|--------------------------------|
| PropertyType | <input type="text" value="0"/> | NBalconies | <input type="text" value="0"/> |
| Country | <input type="text" value="0"/> | PropertyAge | <input type="text" value="0"/> |
| Governorate | <input type="text" value="0"/> | RentalPeriod | <input type="text" value="0"/> |
| City | <input type="text" value="0"/> | Contract | <input type="text" value="0"/> |
| Neighborh... | <input type="text" value="0"/> | PropertyCo... | <input type="text" value="0"/> |
| MountainM... | <input type="text" value="0"/> | Side | <input type="text" value="0"/> |
| NRooms | <input type="text" value="0"/> | Residential... | <input type="text" value="0"/> |
| NBathrooms | <input type="text" value="0"/> | AccessRoad | <input type="text" value="0"/> |
| NKitchens | <input type="text" value="0"/> | Population... | <input type="text" value="0"/> |
| NHalls | <input type="text" value="0"/> | Services | <input type="text" value="0"/> |
| NFloor | <input type="text" value="0"/> | Distance | <input type="text" value="0"/> |
| | | StreetType | <input type="text" value="0"/> |
| Adjacentsi... | <input type="text" value="0"/> | | |
| DeluxeSta... | <input type="text" value="0"/> | | |
| FurnitureN... | <input type="text" value="0"/> | | |
| Area | <input type="text" value="0"/> | | |

Figure 6. 2 input Widegets.

Step 7: Implementation of the prediction process.

In [15]:

```
ypred=model.predict([[PropertyType.value,Country.value,Governorate.value,
City.value,Neighborhood.value,MountainMarine.value,
NRooms.value,NBathrooms.value,NKitchens.value,NHalls.value,
NFloor.value,NBalconies.value,PropertyAge.value,RentalPeriod.value,
Contract.value,PropertyCondition.value,Side.value,ResidentialCommercial.va
lue,
```

```

AccessRoad.value,PopulationDensity.value,Services.value,Distance.value,Str
eetType.value,

Adjacentsides.value,DeluxeStandard.value,FurnitureNonFurnished.value,Area.
value]])

txt=ypred[0]
txt

```

6.3 MLP visualization

The following figure represent the visualization of MLP algorithm in the data.

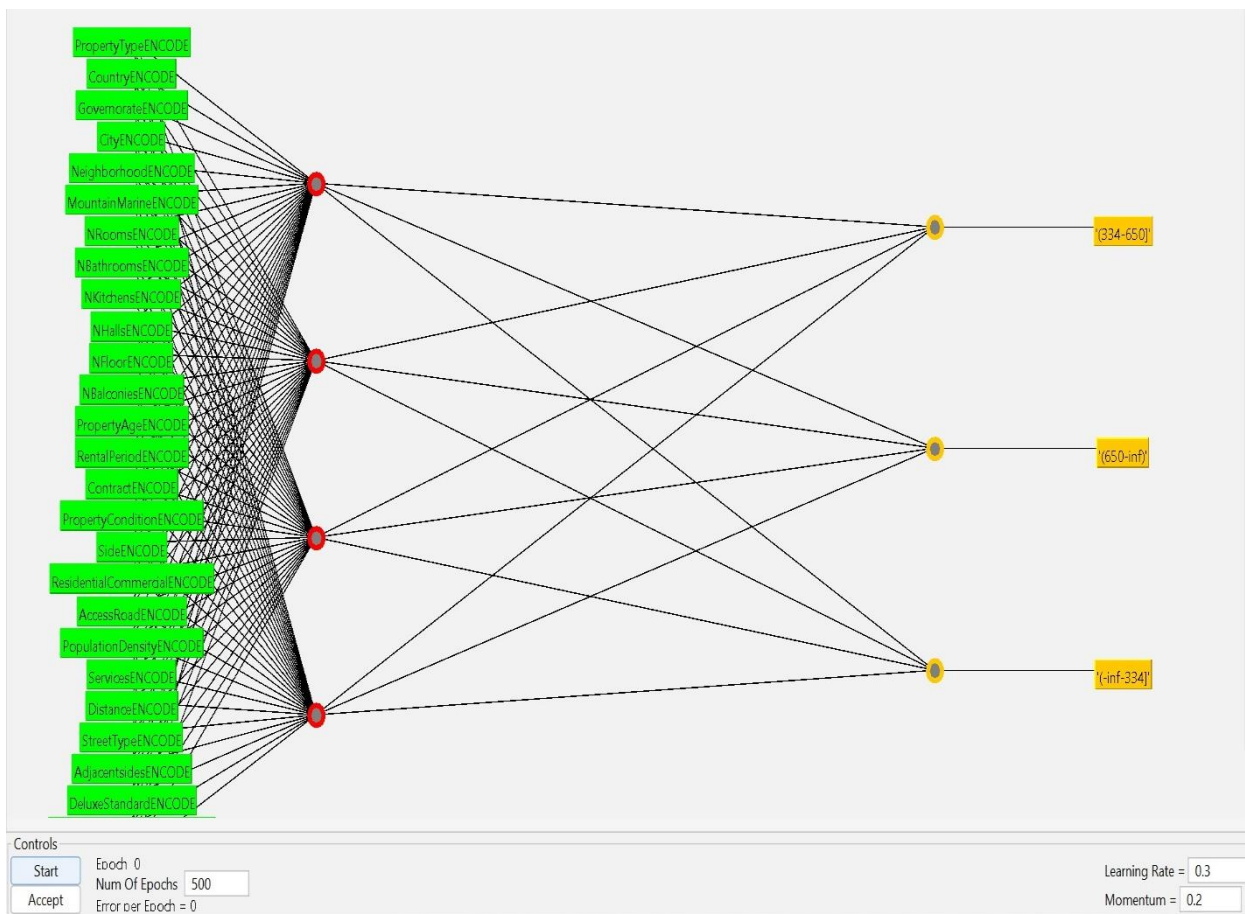


Figure 6. 3 :MLP visualization

The above figure shows the variables connected to the input layers which are shown in the red color and the target variable which is the price has divided into three categories which are 0 to 334, 334 to 650 and 650 to infinite and they are considered as output layer and they are shown in the red color.

Chapter Seven

Conclusion and Future work

7.1 Introduction

This chapter discusses the result that have accomplished in the Rental Estate Advisor system and what will be done in the future work of this system.

7.2 Conclusion

Real estate rental advisor system is a very important system to facilitate and solve many problems that people suffer from in estimating rental property values, and this in turn helps in making critical decisions by giving a clear perception of rental values in many areas.

In this document, the problem of the lack of systems that predict real estate rental prices is discussed in Mukalla-Fouh in Yemen, with many archiving systems that are used to manage property and real estate in general, and they are widespread systems that lack prediction. One of the main contributions of the work is to express this task as a problem, and to suggest ways to solve it based on data science, machine learning, and more precise algorithms chosen and implemented.

7.3 Future work

Many different modifications, tests and experiments were left for the future due to lack of time (i.e. experiments with real data are usually very time consuming, even requiring days to finish part of this work).

Certain mechanisms, new proposals to try different methods, design a website and implement the model in it and there some of the ideas need to develop, such as getting the system to work in several areas, and increase the accuracy of property rental price predictions by expanding the data set.

References

- [1] <https://www.investopedia.com/> ,1/5/2022
- [2] <https://www.weetechsolution.com/> ,1/5/2022
- [3] <https://www.netapp.com/artificial-intelligence/what-is-machine-learning/> ,1/5/2022
- [4] <https://www.ibm.com/cloud/learn/machine-learning> ,5/5/2022
- [5] <https://www.expert.ai/blog/machine-learning-definition/> ,5/5/2022
- [6] <https://www.ibm.com/cloud/learn/data-science-introduction> ,5/5/2022
- [7] <https://www.analytixlabs.co.in/blog/why-do-we-need-data-science/> ,7/5/2022
- [8] <https://dm-consulting.biz/> ,7/5/2022
- [9] <https://ijesc.org/upload/ccf0be7.RealEstate%20House%20Prediction%20usingMachineLearning.pdf> ,7/5/2022
- [10] https://www.researchgate.net/publication/354403038_Bangalore_House_Price_Prediction ,10/5/2022
- [11] https://www.researchgate.net/publication/349477129_House_Price_Prediction ,10/5/2022
- [12] <https://www.amlaaki.com/> ,10/5/2022
- [13] <https://www.amlaaki.com/index.php?p=estethmar> ,10/5/2022
- [14] <http://www.arabic2.com/> ,10/5/2022
- [15] <http://falconproarabic.blogspot.com/> ,15/5/2022s
- [16] <https://fekrait.com> ,15/5/2022
- [17] <http://eol.smartappsye.com/All/posts> ,15/5/2022
- [18] <https://www.finder.com.au/reescienco,a%20property's%20sale%20price> ,15/5/2022
- [19] <https://realas.com/> ,19/5/2022
- [20] <https://www.finder.com.au/realas> ,19/5/2022
- [21] <https://www.investopedia.com/articles/personal-finance/021815/zillow-vs-trulia.asp> ,19/5/2022
- [22] <https://www.computerhope.com/> ,19/5/2022
- [23] <https://towardsdatascience.com/what-is-data-ade94b37204a> ,19/5/2022
- [24] <https://www.talend.com/resources/what-is-data-preparation/> ,19/5/2022
- [25] [https://towardsdatascience.com/missing-value-handling-missing-data-types-](https://towardsdatascience.com/missing-value-handling-missing-data-types-.) ,19/5/2022
- [26] <https://towardsdatascience.com/missing-value-handling-missing-data-typesa89c0d81a5bb#:~:text=There%20are%20four%20types%20of,of%20multiple%20missing%20data%20types> ,19/5/2022

- [27] <https://www.statisticshowto.com/probability-and-statistics/interquartile-range/#IQRExcel>, 19/5/2022
- [28] <https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15> ,27/8/2022
- [29] https://www.investopedia.com/terms/d/descriptive_statistics.asp ,27/8/2022
- [30] <https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization> ,27/8/2022
- [31] <https://sixsigmastudyguide.com/data-distributions/> ,27/8/2022
- [32] <https://www.dominodatalab.com/data-science-dictionary/model-evaluation> , 27/8/2022
- [33] <https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5> ,27/8/2022
- [34] https://www.tutorialspoint.com/weka/what_is_weka.htm ,27/8/2022
- [35] <https://h2o.ai/wiki/multilayer-perceptron/> ,27/8/2022
- [36] <https://deepai.org/machine-learning-glossary-and-terms/multilayer-perceptro> ,27/8/2022

الخلاصة

النماذج التنبؤية لتقرير الإيجار التقديري للعقارات في المدن الكبرى لا تزال مهمة أكثر صعوبة وصعوبة. يعتمد الإيجار المقدر للعقارات على مجموعة متنوعة من العوامل المترابطة. تشمل العوامل الرئيسية ، التي قد تؤثر على الإيجار التقديري للعقار ، مساحة العقار وموقعه ووسائل الراحة فيه. في هذا النظام ، تم إجراء محاولة لبناء نموذج تنبؤي لتقييم الإيجار المقدر بناءً على العوامل التي تؤثر على الإيجار التقديري للعقار. تطبق دراسة النمذجة بعض تقنيات التعلم تحت الإشراف مثل مصنف بايزي أو خوارزميات KNN. تُستخدم هذه النماذج لبناء نموذج تنبؤي ، واختيار أفضل نموذج أداء عن طريق إجراء تحليل مقارن للأخطاء التنبؤية التي تم الحصول عليها بين هذه النماذج. هنا ، تتمثل المحاولة في بناء نموذج تنبؤي لتقييم الإيجار المقدر بناءً على العوامل التي تؤثر على الإيجار التقديري للعقار. تم بناء هذا المفهوم كتطبيق في الوقت الفعلي مفيد للأعمال العقارية وكذلك المشترين والبائعين.