adani | Institute of Digital Technology Management

Capstone Project Report

On

Analyzing and Predicting Real Estate Trends and Location, Using real data.

By

Mr. Nishad Ravindra Kolte

EPBA 02 2023-24

In Partial Fulfillment of the Requirements for the

Executive Program in Business Analytics (EPBA)

at

Adani Institute of Digital Technology Management, Gandhinagar

&

Carleton University, Ottawa, Canada

TABLE OF CONTENTS

Abstract

Understanding how well a country is doing economically involves looking at many different signs, like property prices. Studying these signs has changed how people write about the economy in many ways. There aren't a lot of studies on this topic, but that's a reason to do more research. It shows there's a part of the economy that hasn't been explored much, and it doesn't take a lot of resources to learn more about it, which could open up new business and academic opportunities.

This study is dedicated to addressing the challenge of accurately estimating property prices, focusing specifically on apartment prices in Different cities of Gujarat and Madhya Pradesh. We wanted to shine a light on how data analysis can help with this, using different techniques like processing data, analyzing it, and making predictions. We looked at different budgets suggested by people according to the cities and where they're located to figure out apartment prices.

Our goal is to add useful information to the existing knowledge by using Big Data Analysis, Power BI software, and machine learning techniques. We're especially looking at how these techniques work and what they need in terms of conditions, different parts, and requirements.

Problem Statement

The current challenge in the real estate sector lies in the effective analysis and interpretation of vast datasets, encompassing property details, market trends, and geographical information. As the volume of real estate data continues to grow, there is a pressing need to address key issues such as inaccurate property valuation, inefficient market predictions, and a lack of comprehensive insights into emerging trends. Additionally, the diverse and often unstructured nature of real estate data poses a significant obstacle to extracting meaningful conclusions. This problem statement underscores the imperative to develop robust data analysis methodologies that can enhance the accuracy of property valuations, improve market predictions, and provide valuable insights for stakeholders in the dynamic and ever-evolving real estate landscape.

Introduction

In the ever-evolving landscape of the real estate market, the ability to glean meaningful insights from data has become paramount for informed decision-making. This data analysis project embarks on using a mix of tools, like cleaning up data, creating visual graphs with Power BI, predicting future trends, using EViews software, and making sure all the information stays safe.

Our visualization techniques will not only provide a clear and accessible representation of the data but will also serve as a crucial tool for stakeholders to grasp complex relationships and trends intuitively. Time series forecasting will be employed to project future market trends, allowing for proactive decision-making based on anticipated shifts in the real estate landscape.

As we delve into the factors influencing the success of property shows, machine learning models will contribute valuable predictive insights.

Finally, a paramount consideration throughout our project is data security. We recognize the sensitivity of the information involved and implement stringent measures to safeguard against unauthorized access or data breaches. By prioritizing data security, we ensure the confidentiality and privacy of the information under scrutiny.

In a nutshell, our project is all about understanding what makes property shows successful. We're using different tools and techniques to dig into the details, so we can share helpful insights that can make property shows even better in the future.

Project Task

1. **Data Cleaning and data summarization:** Initially we cleaned and pre-process the dataset, including removing missing values, dealing with outliers, and handling duplicate entries. Also, we have added a few columns and distributed the following data according to the header, further the budget ranges are categorized as provided in dataset.
2. **Data Visualization:** We have used Power BI to create visualizations to explore the dataset, including Pie charts and others to identify patterns and trends in the data. Also prepared a dashboard exhibiting good visualization and data story telling.
3. **Time Series Forecasting:** We have build a time series model to forecast trends based on data. Evaluating the model using metrics such as accuracy and precision.
4. **Machine Learning:** Discussed in theoretical part of some key topics like Supervised learning, Unsupervised learning, and Deep learning.
5. **Data Security:** Excel sheets can be used to demonstrate data security by implementing various security measures to protect sensitive data from unauthorized access, modification, or deletion. Here are some ways in which you may implement in Excel to demonstrate data security: Encryption, Data validation, Restrict access, Audit Trail and Recovery.

Detailed Analysis

1. Power Bi Dashboard

Utilizing the dataset obtained from a real estate expo in Gujarat posed challenges in discerning the precise budgetary constraints of individuals within a specific location. Consequently, we undertook the task of categorizing the data based on property size/type and location. Each registration in the dataset is linked to a unique individual who has indicated their budgetary preferences in correlation to the desired property type within a specific vicinity.

Our Power BI dashboard incorporates an array of visuals, including Tree maps, Filled Maps, Pie Charts, Slicer Cards, and a Table. To provide a more comprehensive understanding, let's delve into each component of the dashboard, elucidating their significance and the insights they yield.

- **Tree maps:**

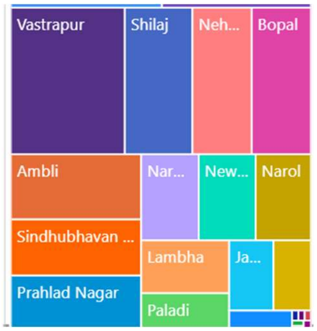


Figure (a)

In the Figure (a) above, we have seamlessly interconnected various locations mentioned in the dataset with the inquiries made by individuals. To facilitate precise data retrieval for each location, we have employed each location tab as a clickable button. For instance, selecting the "Narol" tab triggers a dynamic response wherein all other visuals in the dashboard exclusively display data pertaining to Narol. This interactive feature proves invaluable in swiftly pinpointing the specific information sought by our clients.

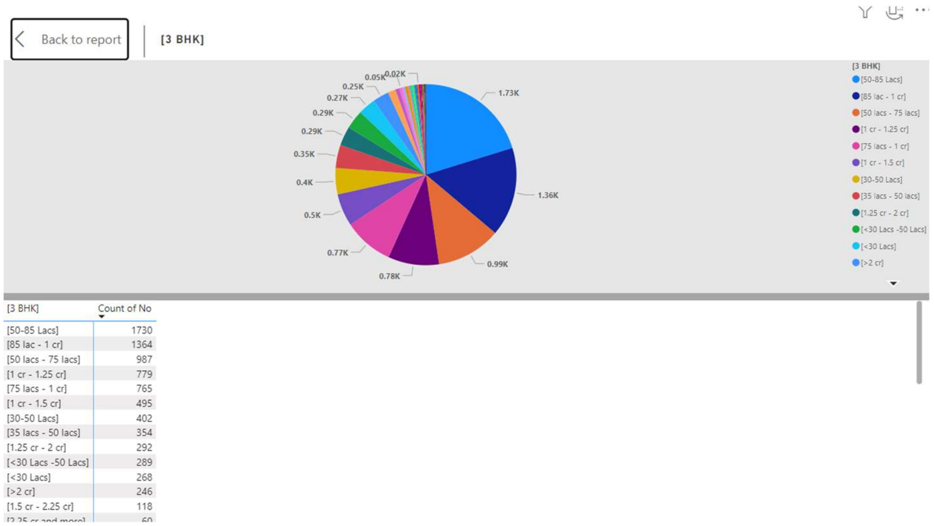
- **Filled Maps:**

Location
PRANTIYA
ANDESAN
RATANPU
PIROJPUR
Location Gift City
© 2023 Microsoft Corporation Terms
Sabarmati
VAL

Figure(b)

As outlined in the preceding passage, the utilization of Filled Maps serves as a pivotal tool in rendering a visual representation of the designated red area, indicating the distinct regions where real estate visitors have expressed keen interest in acquiring both new and existing properties.

- **Pie Chart:**



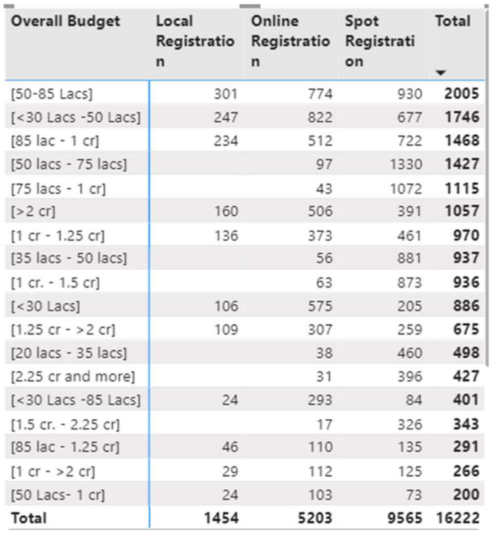
[3 BHK]	Count of No
[50-85 Lacs]	1730
[85 lac - 1 cr]	1364
[50 lacs - 75 lacs]	987
[1 cr - 1.25 cr]	779
[75 lacs - 1 cr]	765
[1 cr - 1.5 cr]	495
[30-50 Lacs]	402
[35 lacs - 50 lacs]	354
[1.25 cr - 2 cr]	292
[<30 Lacs -50 Lacs]	289
[<30 Lacs]	268
[>2 cr]	246
[1.5 cr - 2.25 cr]	118
[2.25 cr and more]	60

Figure(c)

In Figure (c), the Pie Chart illustrates the count of expo visitors expressing interest in 1 BHK properties, categorized based on their budgetary allocations. Nine distinct pie charts have been generated, each offering precise information tailored to the specific property type and location. For a comprehensive analysis, please refer to the detailed insights provided in the key findings section.

In summary, each visual component in our Power BI dashboard plays a crucial role in unraveling the intricate patterns within the real estate dataset, empowering users to make informed decisions by gaining valuable insights into budgetary trends and property preferences across different dimensions.

- **Table Chart:**

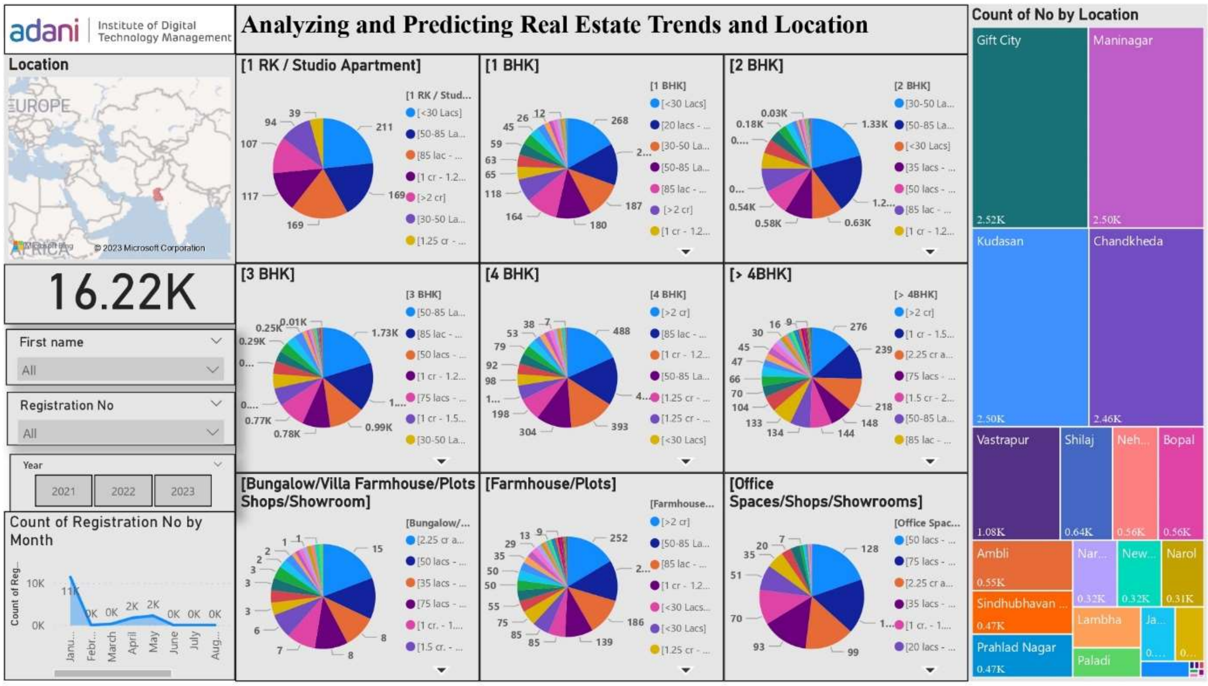


Overall Budget	Local Registration	Online Registration	Spot Registration	Total
[50-85 Lacs]	301	774	930	**2005**
[<30 Lacs -50 Lacs]	247	822	677	**1746**
[85 lac - 1 cr]	234	512	722	**1468**
[50 lacs - 75 lacs]		97	1330	**1427**
[75 lacs - 1 cr]		43	1072	**1115**
[>2 cr]	160	506	391	**1057**
[1 cr - 1.25 cr]	136	373	461	**970**
[35 lacs - 50 lacs]		56	881	**937**
[1 cr. - 1.5 cr]		63	873	**936**
[<30 Lacs]	106	575	205	**886**
[1.25 cr - >2 cr]	109	307	259	**675**
[20 lacs - 35 lacs]		38	460	**498**
[2.25 cr and more]		31	396	**427**
[<30 Lacs -85 Lacs]	24	293	84	**401**
[1.5 cr. - 2.25 cr]		17	326	**343**
[85 lac - 1.25 cr]	46	110	135	**291**
[1 cr - >2 cr]	29	112	125	**266**
[50 Lacs- 1 cr]	24	103	73	**200**
Total	**1454**	**5203**	**9565**	**16222**

Figure (d)

In the depicted figure (d), we present a tabular chart displaying the registration statistics, categorizing them by type—whether conducted online, on the spot, or locally. Each registered visitor has furnished their comprehensive budget information. Consequently, we have organized the entries based on their respective budget categories.

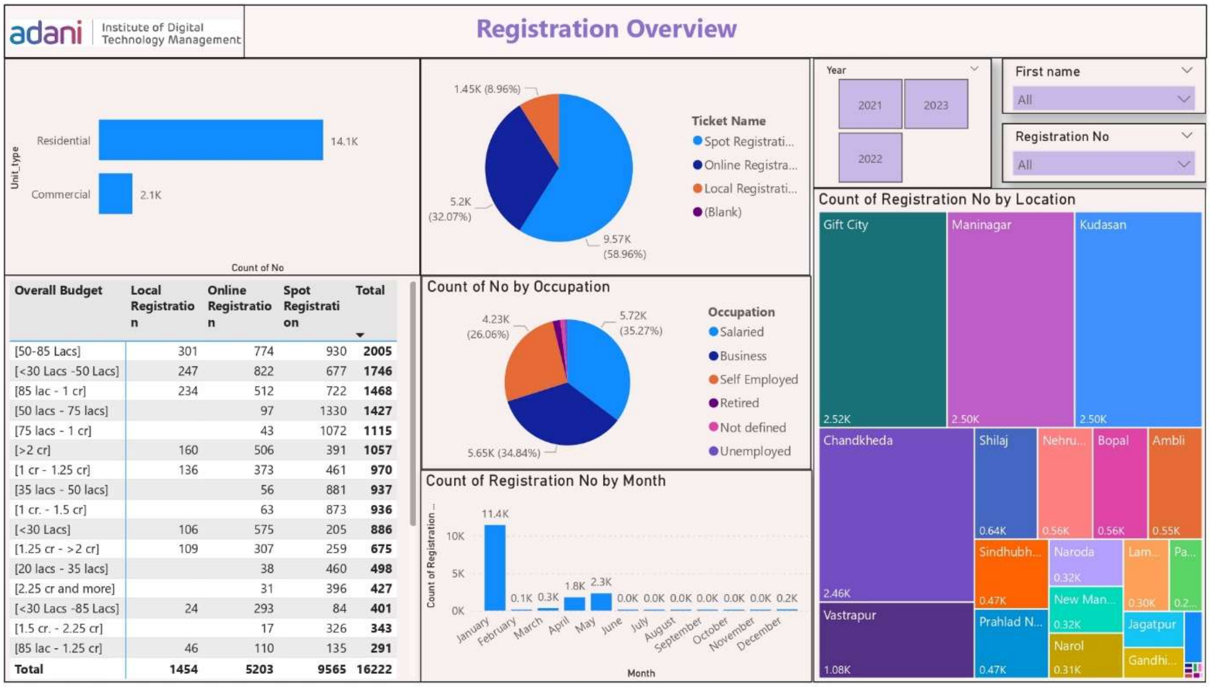
In summary, the presented figure goes beyond a mere enumeration of registration numbers. It serves as a dynamic visual storytelling tool, unraveling the intricate web of registration dynamics and providing a nuanced understanding of the diverse factors influencing attendee participation.



Analyzing and Predicting Real Estate Trends and Location

16.22K

Power Bi Report Page 1



Registration Overview

Overall Budget	Local Registration	Online Registration	Spot Registration	Total
[50-85 Lacs]	301	774	930	**2005**
[<30 Lacs -50 Lacs]	247	822	677	**1746**
[85 lac - 1 cr]	234	512	722	**1468**
[50 lacs - 75 lacs]		97	1330	**1427**
[75 lacs - 1 cr]		43	1072	**1115**
[>2 cr]	160	506	391	**1057**
[1 cr - 1.25 cr]	136	373	461	**970**
[35 lacs - 50 lacs]		56	881	**937**
[1 cr. - 1.5 cr]		63	873	**936**
[<30 Lacs]	106	575	205	**886**
[1.25 cr - >2 cr]	109	307	259	**675**
[20 lacs - 35 lacs]		38	460	**498**
[2.25 cr and more]		31	396	**427**
[<30 Lacs -85 Lacs]	24	293	84	**401**
[1.5 cr. - 2.25 cr]		17	326	**343**
[85 lac - 1.25 cr]	46	110	135	**291**
Total	**1454**	**5203**	**9565**	**16222**

Power Bi report Page

2. Time series Model

Introduction

The dataset, derived from a live real estate expo held in Ahmedabad, encapsulates valuable information regarding the event's registration metrics. It comprises not only the aggregate count of registrations but also delves into the individual visitor profiles, encompassing personal details such as names, email addresses, and mobile numbers. Furthermore, the dataset meticulously captures the specific property preferences of each attendee, including their desired property type and preferred location. Notably, every participant has articulated distinct budgetary considerations tailored to their envisioned property. This comprehensive dataset thus provides a granular perspective on the participants' engagement with the real estate offerings, offering a rich reservoir of insights into the nuanced dynamics of property preferences and financial considerations.

Data Collection and Preprocessing:

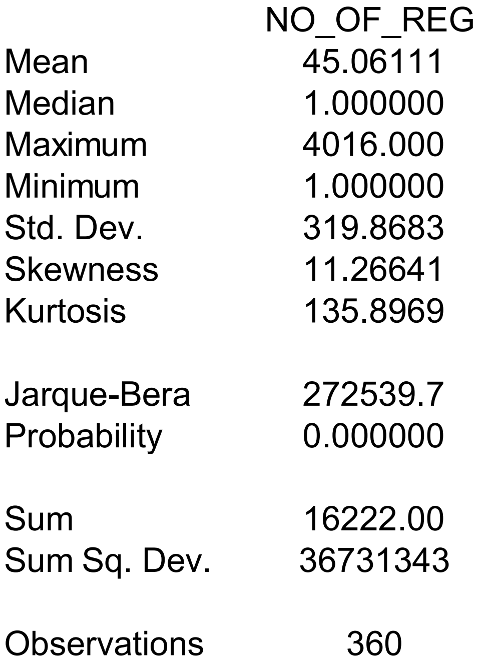
The dataset encompasses a diverse spectrum of budgetary allocations, exemplified by ranges such as [<30 lakhs – 50 lakhs] designated for 1 BHK residences in Gift City. Analogously, a variety of budget ranges are discernible within the dataset, correlating with different property types across distinct locations. It is imperative to acknowledge that these ranges, while informative, are not suitable for the development of a predictive model.

Model Selection:

We have selected the count of registrations per month as the dependent variable and developed a time series model to forecast the registration figures for the upcoming month. This forecasting relies on the historical registration data from preceding months, serving as the basis for predicting future registration trends.

In below figure (e) of descriptive statistic we have observed Probability is > 0.1 that clearly indicates the Data is not normally distributed

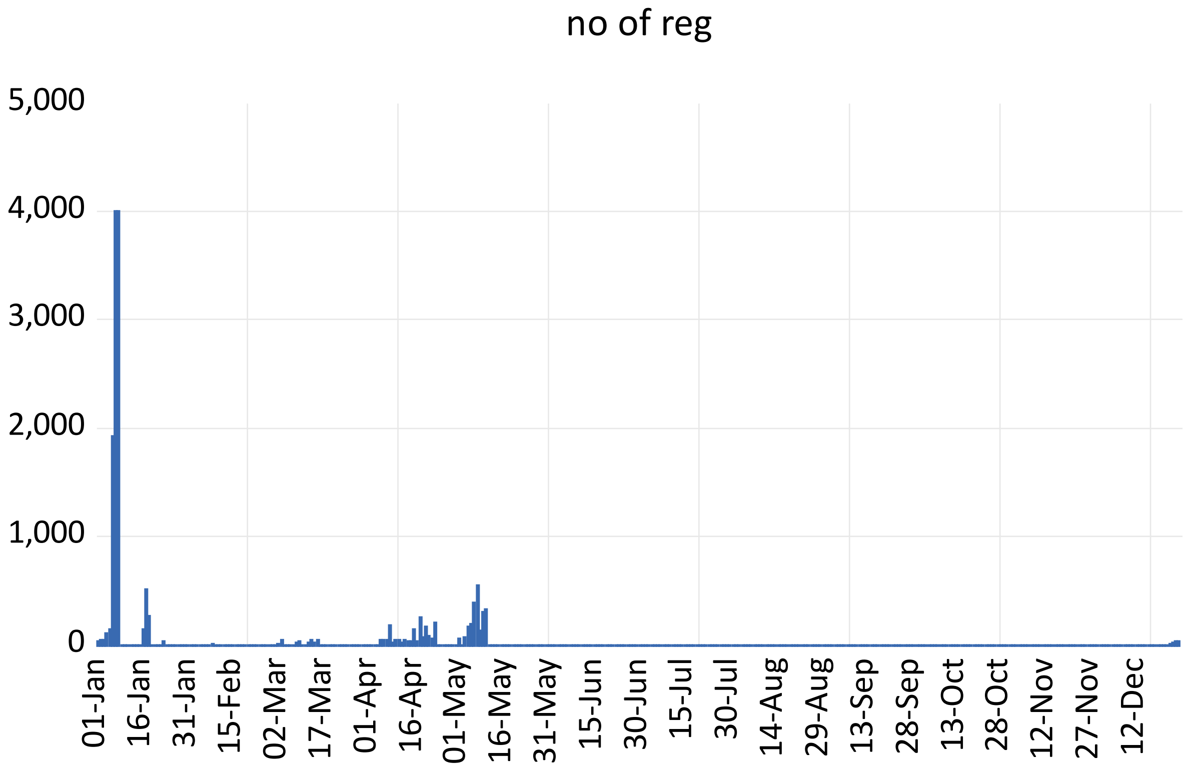
Descriptive Statistics model:



```
                    NO_OF_REG
Mean                 45.06111
Median               1.000000
Maximum              4016.000
Minimum              1.000000
Std. Dev.            319.8683
Skewness             11.26641
Kurtosis             135.8969

Jarque-Bera          272539.7
Probability          0.000000

Sum                  16222.00
Sum Sq. Dev.         36731343

Observations            360
```

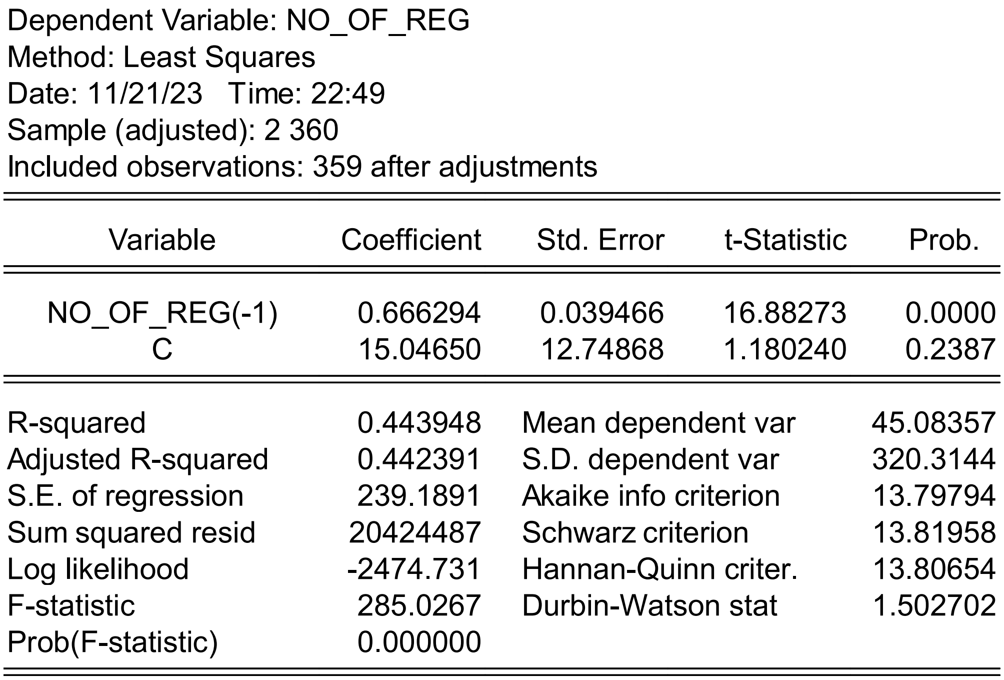
Figure (e)

Bell curve :



For more clarity please refer the above chart that indicates the data is not normally distributed, the bell curve formation is on the left side as the data is accumulated towards left.

Time Series Model :



```
Dependent Variable: NO_OF_REG
Method: Least Squares
Date: 11/21/23   Time: 22:49
Sample (adjusted): 2 360
Included observations: 359 after adjustments
==========================================================================
       Variable          Coefficient    Std. Error    t-Statistic    Prob.
==========================================================================
    NO_OF_REG(-1)          0.666294      0.039466      16.88273     0.0000
          C                15.04650      12.74868      1.180240     0.2387
==========================================================================
R-squared                  0.443948   Mean dependent var          45.08357
Adjusted R-squared         0.442391   S.D. dependent var          320.3144
S.E. of regression         239.1891   Akaike info criterion       13.79794
Sum squared resid          20424487   Schwarz criterion           13.81958
Log likelihood            -2474.731   Hannan-Quinn criter.        13.80654
F-statistic                285.0267   Durbin-Watson stat          1.502702
Prob(F-statistic)          0.000000
==========================================================================
```

Above Time series model is using Least square method, the dependent variable is No of registration per month of the dataset

B_0 = Constant

B_1 = No of Registration

Equation : $Y = B_0 + B_1 (x) + \text{Error}$

No of registrations is dependent variable, being predicted or explained by the model. The independent variables are no_of_reg(-1): 0.666294 which is lagged value of dependent variable and constant C: 15.04650. The model suggests that the current value of no. of registrations is influenced by its past value and it is indicated by the coefficient of 0.666294 with the low p value as 0.0000.

R-squared is 0.443948, this explains about 44% of the variability in the number of registrations. Durbin Watson statistic (1.502702) is used to test of autocorrelation. a value approximately equal to 2 suggests that there's no auto correlation.

Therefore, the number of registrations are dependent variable in the data which is dependent on the past month registration

Machine Learning

- **Supervised learning**

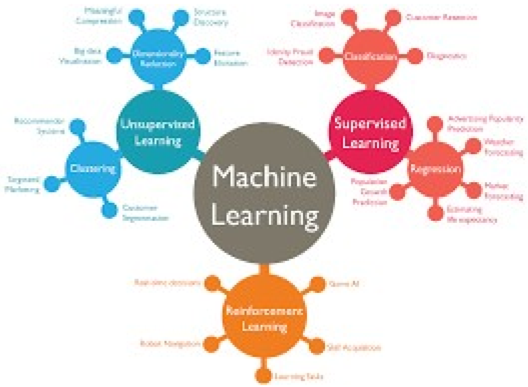
Linear regression

Linear regression is a fundamental technique in machine learning and statistics used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. In simple terms, it helps us understand the linear relationship between input features and the output.

y=mx+b

Decision tree

A decision tree is a flowchart-like structure used in machine learning and statistics to make decisions. It breaks down a decision-making process into a series of choices and their possible outcomes. Each node in the tree represents a decision or a test on a specific feature, and each branch represents the outcome of that decision or test. It's a powerful tool for classification and regression tasks.



- **Unsupervised learning**

Clustering

Clustering is a technique in unsupervised learning where data points are grouped together based on similarities, aiming to discover inherent patterns or structures within the data. Popular algorithms for clustering include K-means, hierarchical clustering.

K means clustering

The goal of K-means is to partition data points into K clusters based on similarity, with each cluster represented by its centroid (mean).

Applications:

- Customer segmentation
- Image compression
- Document clustering
- Anomaly detection

Neural network

Neural networks are a fundamental concept in machine learning, specifically in the subfield of deep learning.

Neural Network Applications in Machine Learning:

- Image Recognition
- Natural Language Processing (NLP)
- Speech Recognition
- Gaming
- Robotics

Computer vision

Computer vision is a field of artificial intelligence (AI) and computer science that focuses on enabling machines to interpret, understand, and make decisions based on visual data. It aims to replicate the human visual system and perception capabilities using computational models and algorithms. Computer vision applications cover a wide range of tasks related to image and video analysis.

Data Security

Encryption: Encryption is a security technique that transforms information into an unreadable format using algorithms. This process ensures that only authorized parties can access and decipher the data, providing a robust layer of protection against unauthorized access or data breaches. Encryption is essential for safeguarding sensitive information during transmission or storage, adding a crucial element to data security in various contexts such as communication, financial transactions, and data storage.

Data Validation: Data validation is a critical step in ensuring the accuracy and integrity of information within a system. It involves the examination and verification of data to confirm that it complies with predefined standards or rules. By implementing data validation measures, organizations can prevent errors, inconsistencies, and potential security vulnerabilities arising from inaccurate or malicious data inputs. This process is fundamental for maintaining data quality and reliability, contributing to the overall trustworthiness of systems and applications.

Restrict Access: Restricting access is a fundamental security practice that involves controlling and limiting the permissions granted to users, applications, or devices. By defining and enforcing access controls, organizations can mitigate the risk of unauthorized access to sensitive information. Restricting access ensures that only individuals with the necessary credentials or permissions can view, modify, or interact with specific data or system resources. This principle is a cornerstone of information security and is essential for protecting confidential and valuable assets.

Audit Trail: An audit trail is a chronological record of activities, events, or transactions within a system or application. It serves as a detailed and traceable history, documenting who accessed what information, when, and what actions were taken. Audit trails are crucial for monitoring and analyzing system activity, identifying security incidents, and facilitating compliance with regulatory requirements. By maintaining comprehensive audit trails, organizations can enhance their ability to detect and respond to security threats, investigate incidents, and demonstrate accountability.

Recovery: Recovery in the context of information systems refers to the process of restoring data, services, or systems to a functional state after an incident or disaster. This may involve recovering data from backups, rebuilding infrastructure, or implementing contingency plans. A robust recovery strategy is essential for minimizing downtime, mitigating the impact of disruptions, and ensuring business continuity. It encompasses backup procedures, disaster recovery plans, and the ability to recover critical systems and data efficiently, thus enhancing the overall resilience of an organization in the face of unforeseen events.

Key Findings/Deliverables

Sr no	Type of property	Location	Budget (INR)	Majority Registration
1	1 RK / Studio Apartment	Naroda, New maninager, Narol	< 30 Lacs	211 Nos
2	1 BHK	Lambha, Naroda, New maninager, Narol	< 30 Lacs	268 Nos
3	2 BHK	Gift city, maninagar,Kudasan,chandkheda,	30-50 Lacs	1333 Nos
4	3 BHK	Gift city, maninagar,Kudasan,chandkheda	50-85Lacs	1730 Nos
5	4 BHK	Vastrapur,Shilaj, sidhubhavan,Prahlad nagar	>2 Cr	818 Nos
6	>4 BHK	Vastrapur,Shilaj, sidhubhavan,Prahlad nagar	2.25 Cr and more	713 Nos
7	Bungalow/Villa	Shilaj,	2.25 Cr and more	25 Nos
8	Farmhouse	Shilaj, Sindhubhavan, Vastrapur	>2 Cr	252 Nos
9	Office spaces	Kudsan,Gift City	50-75Lacs	128 Nos

The data above highlights a common trend among visitors, indicating shared preferences based on property features. In essence, this insight suggests that a builder can tailor property construction to align with the prevailing preferences in a specific location, thereby enhancing sales prospects

Conclusion

In conclusion, this comprehensive data analysis project has been instrumental in transforming raw real estate data into meaningful insights, leveraging a multi-faceted approach that spans data cleaning, Power BI visualization, time series modeling with machine learning, and robust data security practices.

The initial phase of data cleaning was pivotal in ensuring the accuracy and reliability of our dataset. By addressing missing values, outliers, and inconsistencies, we laid a solid foundation for subsequent analysis, enhancing the overall quality of the insights derived from the data.

The integration of Power BI visualization brought the data to life, providing stakeholders with a dynamic and intuitive platform to interact with key metrics and trends. The visualizations not only facilitated a clearer understanding of the real estate landscape but also enabled more informed decision-making by presenting complex information in a accessible manner.

The implementation of a time series model using machine learning algorithms added a predictive dimension to our analysis. By forecasting real estate trends over time, we equipped decision-makers with valuable foresight, allowing them to anticipate market dynamics, identify potential opportunities, and proactively address challenges.

Crucially, throughout every phase of the project, data security remained a top priority. Encryption measures were implemented to safeguard sensitive real estate information, ensuring that only authorized personnel could access and analyze the data. Access controls, audit trails, and other security protocols were meticulously employed to protect the confidentiality and integrity of the data throughout the analysis lifecycle.

As we conclude this project, we not only offer actionable insights for strategic decision-making in the real estate sector but also emphasize the importance of maintaining data security and integrity in an era where data-driven decisions are paramount. This project stands as a testament to the power of a holistic data analysis approach, where cleanliness, visualization, predictive modeling, and security measures collectively contribute to unlocking the full potential of real estate data.