VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belagavi – 590 018

An Internship Report on

"HOUSING PRICES"

Submitted in partial fulfillment for the award of degree of

Bachelor of Engineering

in

Computer Science Engineering

Submitted By

Daniel D (4AD18CS017)

Mohammed Anas (4AD18CS040)

Internship Carried Out at

AUDAZ VENTURES PVT. LTD.

ATME College of Engineering

INTERNAL GUIDE

ATME College of Engineering
Mysuru

AUDAZ

EXTERNAL GUIDE

Mr. Vikrant Kumar
L&D Manager
AUDAZ Venture PVT. LTD.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ATME COLLEGE OF ENGINEERING

13th KM Stone, Mysuru – Bannur Road, Mysuru – 570028

Phone: 0821 2954081

Website: www.atme.in

2022 – 2023

ATME College of Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the internship report titled **"HOUSING PRICES"** is carried out by **Daniel D** bearing USN **4AD18CS017** and **Mohammed Anas** bearing USN **4AD18CS040** in partial fulfillment of the requirements for the award of **Bachelor of Engineering in Computer Science and Engineering** of **Visvesvaraya Technological University, Belagavi** during the year **2022-2023**.

Signature of the Internal Guide	**Signature of HOD**	**Signature of the External Guide**
Dept. of CSE ATME	**Dr. Puttegowda** Professor & HOD Dept. of CSE ATME	**Mr. Vikrant Kumar** L&D Manager AUDAZ Ventures Private Limited

DECLARATION

I hereby declare that I have completed my four weeks Internship at **"AUDAZ VENTURES PVT. LTD.",** from 21-08-2022 to 17-09-2022 under the guidance of the internal guide. I have declared that I have worked with full dedication during these four weeks of Internship in partial fulfillment for the award of the degree of **Bachelor of Engineering in Computer Science and Engineering** from **Visvesvaraya Technological University, Belagavi** during the year **2022 - 23**.

Daniel D (4AD18CS017)

Mohammed Anas (4AD18CS040)

ACKNOWLEDGEMENT

The success and outcome of this internship required a lot of guidance and assistance from many people, and I am extremely privileged to have got this all along the completion of my internship. I thank to all those who have rendered their cherished advice and services towards the completion of the internship.

I wish to express my deep sense of acknowledgement and gratitude to my Internal guide Department of Computer Science and Engineering, for the suggestions and encouragement throughout the making of the internship.

I wish to express my deep sense of acknowledgement and gratitude to my External guide **Mr. Vikrant Kumar,** L&D Manager, Audaz Ventures Private Limited, for the suggestions and encouragement throughout the making of the internship.

I wish to express my deep sense of acknowledgement and gratitude to my Trainers **Mr. Azhar**, **Mr. Mohammed Nawfal** and **Mr. Mohammed Sadiq** for their training and constant support.

I am highly indebted to **Dr. Puttegowda**, Head of Department, Computer science and Engineering, for his kind consents and wholehearted cooperation.

I would like to thank our Principal **Dr. L Basavaraj**, for his encouragement and providing an excellent working environment.

I thank all the lecturers of the dept. for their cooperation and providing with the facilities to carry out the seminar work. I would also express my thanks to all technical and non-technical staff of the Computer Science and Engineering department who have directly or indirectly cooperated with me.

Finally, I would like to express my gratitude to my parents and friends who always stood by me encouraging in all my endeavors.

Daniel D (4AD18CS017)

Mohammed Anas (4AD18CS040)

INDEX

EXECUTIVE SUMMARY

The internship report highlights the major works carried out in terms of academic and non-academic perspectives. The scope of this document is to identify and describe the analysis carried out, project completed, experience gained and focuses on the achievements as an intern.

I was working with AUDAZ VENTURES PVT. LTD. to complete my internship. I found myself rather lucky by getting the chance to work in such an environment that AUDAZ provided and got introduced to some of the new terms, new technologies and languages.

The project that I worked in certainly helped me by increasing my practical knowledge depth. The research and analysis projects were particularly helpful in widening my views regarding data analysis tools and data science. Besides there were some vital lessons which will obviously help in future jobs.

This report intended to describe the analysis of the California Housing Prices. The analysis provides the inferences about the prices of the houses based on latitude and longitude of the city which can be utilized for business efficiencies.

The Chapters One to Six provide the introduction to the project and data analytics. The Chapter Seven explains the data used and data processing for analysis. Finally, in the Chapter Eight to ten the analysis and conclusions based on the project analysis.

INTRODUCTION

As each sector of the market is growing, data is building up day by day, we need to keep the record of the data which can be helpful for the analytics and evaluation. Now we don't have data in gigabyte or terabyte but in petabyte and zettabyte and this data cannot be handled with the day-to-day software such as Excel or MATLAB. Therefore, in this report we will be dealing with large data sets with the high-level programming language 'Python'.

The main goal of this project is to aggregate and analyze the data collected from the different data sources available on the internet. This project mainly focuses on the usage of the Python programming language and its data analytics libraries in the field of housing prices (real estate) industry. This language has not only it's application in the field of just analyzing the data but also for the prediction of the upcoming scenarios.

The purpose of using this specific language is due to its versatility, vast libraries (Pandas, NumPy, Matplotlib, etc.), speed limitations, and ease of learning. We will be analyzing large housing prices datasets in this project which cannot be easily analyzed in other tools as compared to Python. Python does not have its limitation to only data analytics but also in many other fields such as Artificial intelligence, Machine learning, and many more.

This dataset is a modified version of the California Housing dataset available from the University of Porto. It is obtained from the StatLib repository.

COMPANY PROFILE

Audaz Ventures Private Limited

CIN: U80903DL2020PTC365560

Audaz Ventures Private Limited, with the headquarter in New Delhi and corporate office in Bengaluru was established in 2020. It is a service - based company which provides services and solutions to 62 institutes across India. It has its presence in more than 12 cities across India.

Its services include

1. Software solutions
2. Digital marketing
3. SAP software
4. ERP software
5. Blockchain software
6. Placement related training to the Engineering Graduates and many more
7. Placement opportunities to students

Thanks & Regards

Rahul Oberoi

Director

08882264073
09034444243
info@audazlearning.com
C2C/2-100, JANAKPURI,
NEW DELHI - 110058

OBJECTIVES AND SCOPE OF THE STUDY

4.1 Objectives of the study

- To study the data analytics concepts of data gathering, data cleaning, EDA, and result interpretation.
- To study the real estate rate in California.
- To study the least and most priced property in California.

4.2 Scope of Study

- The aim of this study is to analyze the data generated by a housing price.
- The scope of the study is limited to the data available from Housing prices in California.

THEORETICAL BACKGROUND

5.1 What are Data Analytics?

Companies around the globe generate vast volumes of data daily, in the form of log files, web servers, transactional data, and various customer-related data. In addition to this, social media websites also generate enormous amounts of data. Companies ideally need to use all their generated data to derive value out of it and make impactful business decisions. Data analytics is used to drive this purpose.

Data analytics is the process of exploring and analyzing large datasets to find hidden patterns, unseen trends, discover correlations, and derive valuable insights to make business predictions. It improves the speed and efficiency of your business. Businesses use many modern tools and technologies to perform data analytics.

5.2 Ways to Use Data Analytics

- **Improved Decision Making**

  Data Analytics eliminates guesswork and manual tasks. Be it choosing the right content, planning marketing campaigns, or developing products. Organizations can use the insights they gain from data analytics to make informed decisions. Thus, leading to better outcomes and customer satisfaction.

- **Better Customer Service**

  Data analytics allows you to tailor customer service according to their needs. It also provides personalization and builds stronger relationships with customers. Analyzed data can reveal information about customers' interests, concerns, and more. It helps you give better recommendations for products and services.

- **Efficient Operations**

  With the help of data analytics, you can streamline your processes, save money, and boost production. With an improved understanding of what your audience wants, you spend lesser time creating ads and content that aren't in line with audience's interests.

- **Effective Marketing**

  Data analytics gives you valuable insights into how your campaigns are performing. This helps in fine-tuning them for optimal outcomes. Additionally, you can also find potential customers who are most likely to interact with a campaign and convert into leads.

5.3 Steps Involved in Data Analytics

There are a few steps that are involved in the data analytics lifecycle. Below are the steps that you can take to solve your problems.

- **STEP 1: Identify**

  Identifying or understanding the business problems, defining the organizational goals, and planning a lucrative solution is the first step in the analytics process. Real estate agencies often encounter issues such as predicting the cost of the lands etc.
- **STEP 2: Collection**

  Collection or Data collection, is need to collect transactional business data and customer-related information from the past few years to address the problems your business is facing.
- **STEP 3: Clean**

  Clean or data clean, is the data you collect will often be disorderly, messy, and contain unwanted missing values. Such data is not suitable or relevant for performing data analysis. Hence, you need to clean the data to remove unwanted, redundant, and missing values to make it ready for analysis.
- **STEP 4: Analyze**

  Analyze or Data exploration and analysis, is you gather the right data, the next vital step is to execute exploratory data analysis. It can be used as data visualization and business intelligence tools, data mining techniques, and predictive modeling to analyze, visualize, and predict future outcomes from this data. Applying these methods can tell you the impact and relationship of a certain feature as compared to other variables.
- **STEP 5: Interpret the results**:

  Interpret or Interpret of the results, is the final step is to interpret the results and validate if the outcomes meet your expectations. You can find out hidden patterns and future trends. This helps to gain insights that supports with appropriate data-driven decision making.

5.4 Data Analytics Tools

Here are 7 data analytics tools, including a couple of programming languages that can help you perform analytics better.

- **Python** is an object-oriented open-source programming language. It supports a range of libraries for data manipulation, data visualization, and data modeling.

- **R** is an open-source programming language majorly used for numerical and statistical analysis. It provides a range of libraries for data analysis and visualization.

- **Tableau** is a simplified data visualization and analytics tool. This helps you create a variety of visualizations to present the data interactively, build reports, and dashboards to showcase insights and trends.

- **Power BI** is a business intelligence tool that has an easy ‘drag and drop functionality. It supports multiple data sources with features that visually appeal to data. Power BI supports features that help you ask questions to your data and get immediate insights.

- **QlikView** offers interactive analytics with in-memory storage technology to analyze vast volumes of data and use data discoveries to support decision making. It provides social data discovery and interactive guided analytics. It can manipulate colossal data sets instantly with accuracy.

- **Apache Spark** is an open-source data analytics engine that processes data in real-time and carry out sophisticated analytics using SQL queries and machine learning algorithm.

5.5 Data Analytics Applications

Data analytics is used in almost every sector of business, below are a few of them:

- **Retail** helps retailers understand their customer needs and buying habits to predict trends, recommend new products, and boost their business. They optimize the supply chain, and retail operations at every step of the customer journey.

- **Healthcare** industries analyze patient data to provide lifesaving diagnoses and treatment options. Data analytics help in discovering new drug development methods as well.

- **Manufacturing** sectors can discover new cost-saving opportunities. They can solve complex supply chain issues, labor constraints, and equipment breakdowns.

- **Banking sector** uses analytics to find out probable loan defaulters and customer churn out rate. It also helps in detecting fraudulent transactions immediately.

- **Logistics** companies use data analytics to develop new business models and optimize routes. This, in turn, ensures that the delivery reaches on time in a cost-efficient manner.

RESEARCH METHODOLOGY

This research is a descriptive analytic, in which we are analyzing the real estate prices in California. It is based on primary data and inferences derived from it.

6.1 Types of Research in Data Analytics

- **Predictive Analytics**

  It turns the data into valuable, actionable information. predictive analytics uses data to determine the probable outcome of an event or a likelihood of a situation occurring and holds a variety of statistical techniques from modeling, machine, learning, data mining, and game theory that analyze current and historical facts to make predictions about a future event.

- **Descriptive Analytics**

  It looks at data and analyze past event for insight as to how to approach future events. It looks at the past performance and understands the performance by mining historical data to understand the cause of success or failure in the past. Almost all management reporting such as sales, marketing, operations, and finance uses this type of analysis. The descriptive model quantifies relationships in data in a way that is often used to classify customers or prospects into groups.

- **Prescriptive Analytics**

  It automatically synthesizes big data, mathematical science, business rule, and machine learning to make a prediction and then suggests a decision option to take advantage of the prediction. Prescriptive analytics goes beyond predicting future outcomes by also suggesting action benefit from the predictions and showing the decision maker the implication of each decision option. Prescriptive Analytics not only anticipates what will happen and when to happen but also why it will happen.

- **Diagnostic Analytics**

  It generally uses historical data over other data to answer any question or for the solution of any problem. We try to find any dependency and pattern in the historical data of the particular problem.

ANALYSIS AND INTERPRETATION OF DATA

Before analyzing and visualization we need the raw data, and this raw data can be gathered from different open-source data websites available on the internet. This data will be in raw CSV form, it may be individual subunits of data or the complete dataset containing all the groups of files.

7.1 STEP 1: EXTRACT THE DATA

```
1 import pandas as pd
2 df = pd.read_csv("housing.csv")
3 df
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY
...	...	...	...	...	...	...	...	...	...	...
20635	-121.09	39.48	25.0	1665.0	374.0	845.0	330.0	1.5603	78100.0	INLAND
20636	-121.21	39.49	18.0	697.0	150.0	356.0	114.0	2.5568	77100.0	INLAND
20637	-121.22	39.43	17.0	2254.0	485.0	1007.0	433.0	1.7000	92300.0	INLAND
20638	-121.32	39.43	18.0	1860.0	409.0	741.0	349.0	1.8672	84700.0	INLAND
20639	-121.24	39.37	16.0	2785.0	616.0	1387.0	530.0	2.3886	89400.0	INLAND

20640 rows × 10 columns

Reading the data

- **longitude:** A measure of how far west a house is; a higher value is farther west
- **latitude:** A measure of how far north a house is; a higher value is farther north
- **housingMedianAge:** Median age of a house within a block; a lower number is a newer building
- **totalRooms:** Total number of rooms within a block
- **totalBedrooms:** Total number of bedrooms within a block
- **population:** Total number of people residing within a block
- **households:** Total number of households, group of people residing within home unit, for a block
- **medianIncome:** Median income for households within a block of houses (measured in tens of thousands of US Dollars)
- **medianHouseValue:** Median house value for households within block (measured in US Dollars)
- **oceanProximity:** Location of the house w.r.t ocean/sea

7.2 STEP 2: DATA CLEANING

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

info() method prints information about the Data Frame

The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values)

```
1 df.describe()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000	20640.000000
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816909
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615874
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.000000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000000
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.000000
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000

describe() method returns description of the data in the Data Frame.

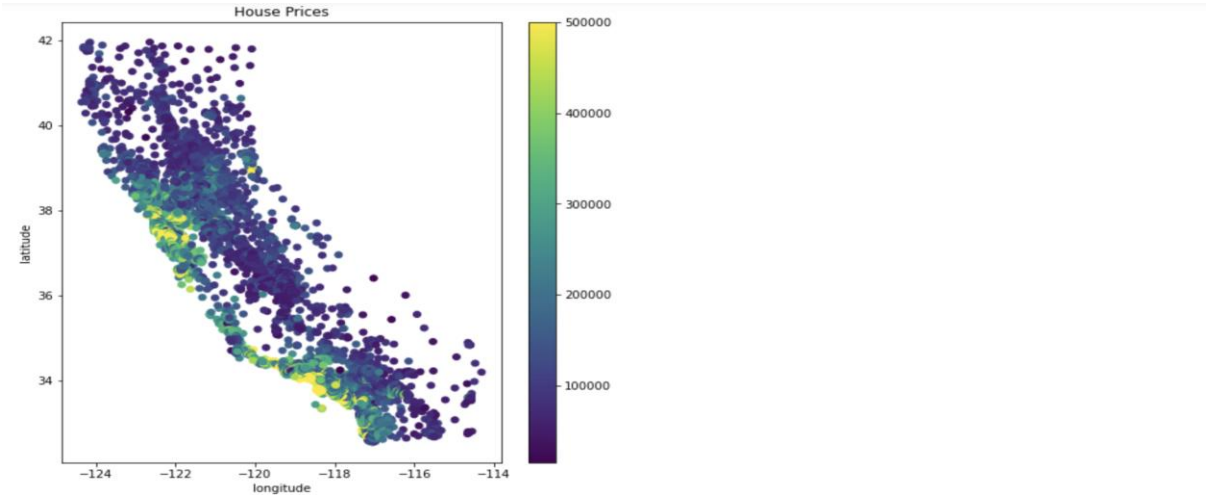
If the Data Frame contains numerical data, the description contains this information for each column: count - The number of not-empty values. mean - The average (mean) value. std - The standard deviation.

7.3 STEP 3: VISUALIZAION OF DATA

SCATTER PLOT

```
# Scatter plot is a type of data visualization that shows the relationship between different variables.
plt.figure(figsize = (8,8))
plt.scatter(df['longitude'] , df['latitude'] , c = df['median_house_value'])
plt.colorbar()
plt.xlabel("longitude")
plt.ylabel("latitude")
plt.title("House Prices")
```

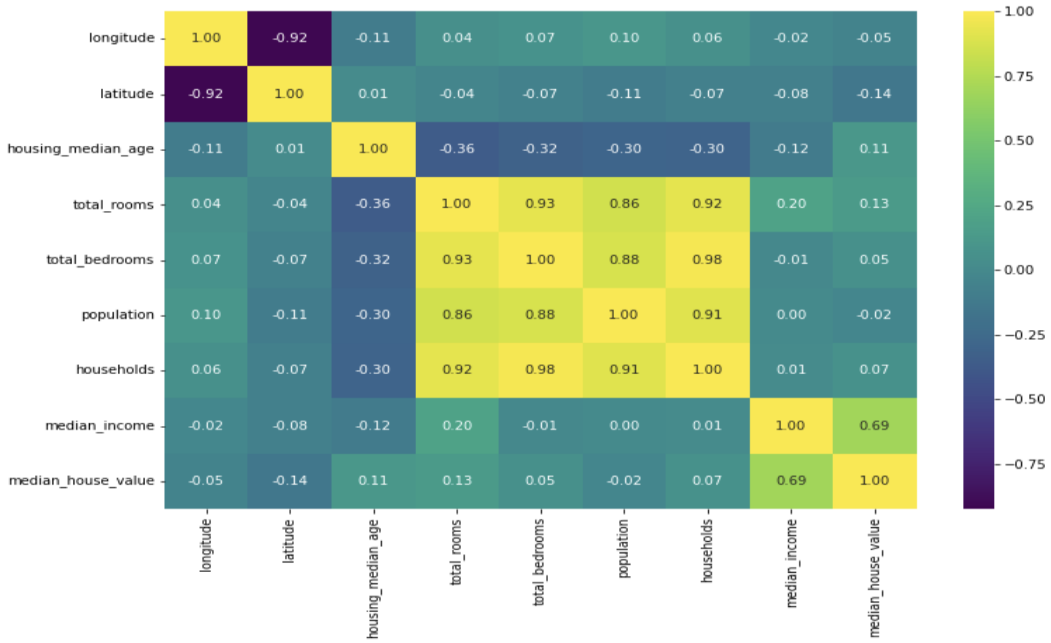
```
Text(0.5, 1.0, 'House Prices')
```



SCATTERPLOT BASED ON LATIRUDE AND LONGITUDE

HEATMAP

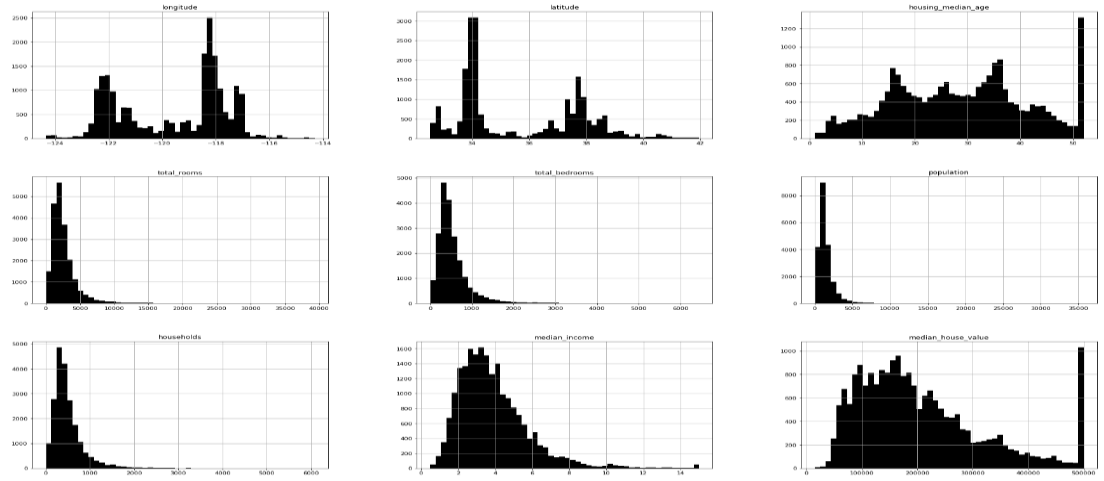
```
# Heatmap is used to visualize visitor behavior data in form of hot and cold spots employing a warm-to-cool color scheme
plt.figure(figsize = (12, 8))
sns.heatmap(df.corr(), annot = True, fmt = '.2f', cmap = 'viridis')
plt.show()
```



Target variable median_house_value is very mildly correlated to all but one feature here median_income, so one might outline this as an important feature.

HISTOGRAM

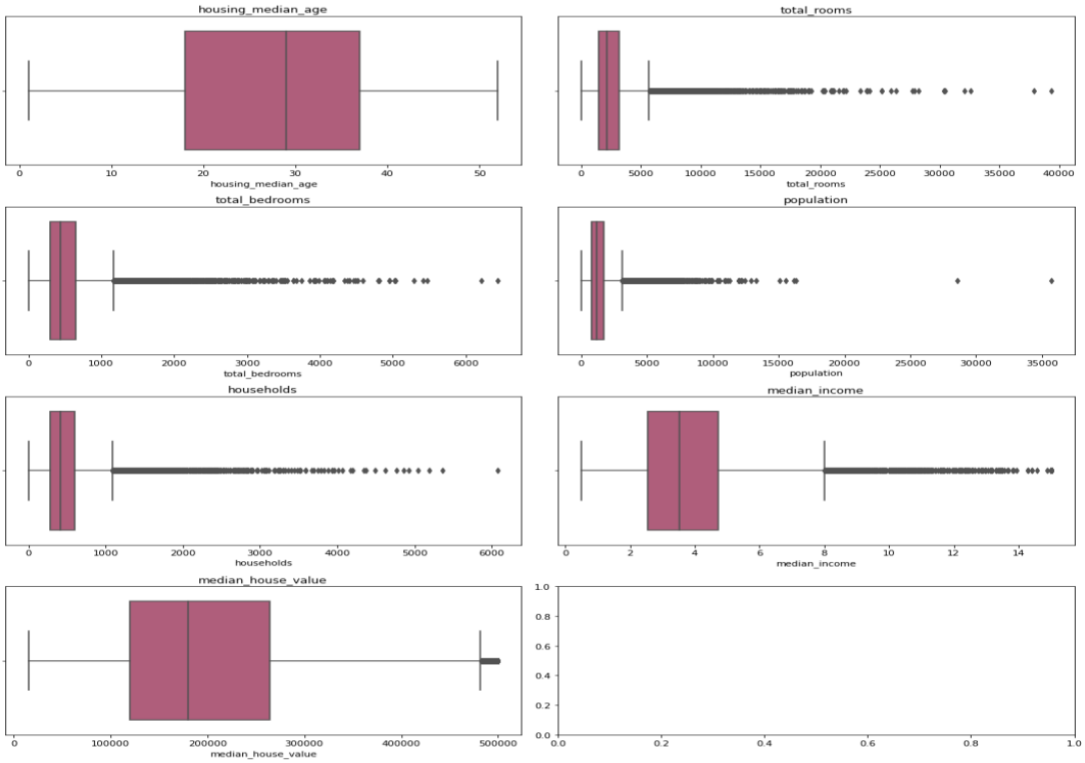
```
# Histogram provides a visual representation of the distribution of a dataset
df.hist(bins = 50 , figsize=(30 , 20),color="k")
plt.show()
```



It is used to observe the data Distribution

BOXPLOT

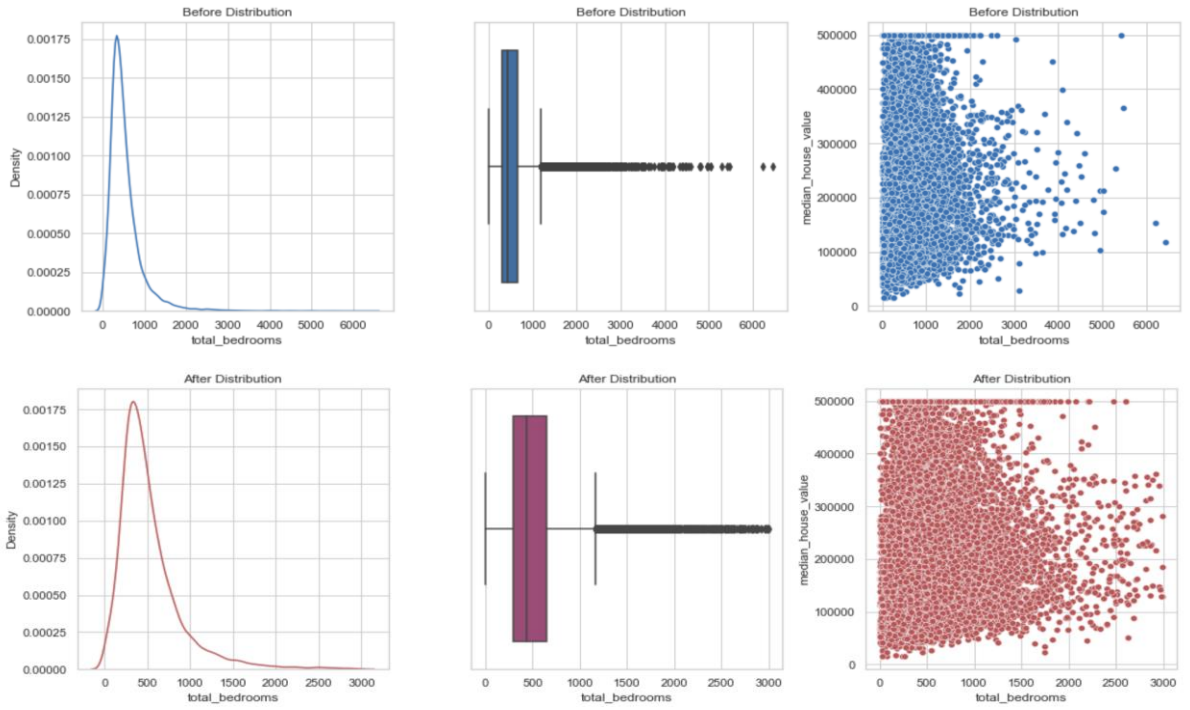
```
# Boxplot allows you to examine the distribution of data
num_columns = list(df.select_dtypes(include=["int64","float64"]).columns)[2:]
fig, ax = plt.subplots(4,2, figsize = (15,15))
font_dict = {'fontsize': 14}
ax = np.ravel(ax)
for i in num_columns:
    sns.boxplot(data=df,x=i,ax=ax[num_columns.index(i)],palette="plasma").set_title(i)
ax = np.reshape(ax, (4, 2))
plt.tight_layout()
plt.show()
```



It is used to examine the data Distribution

DISTRIBUTION

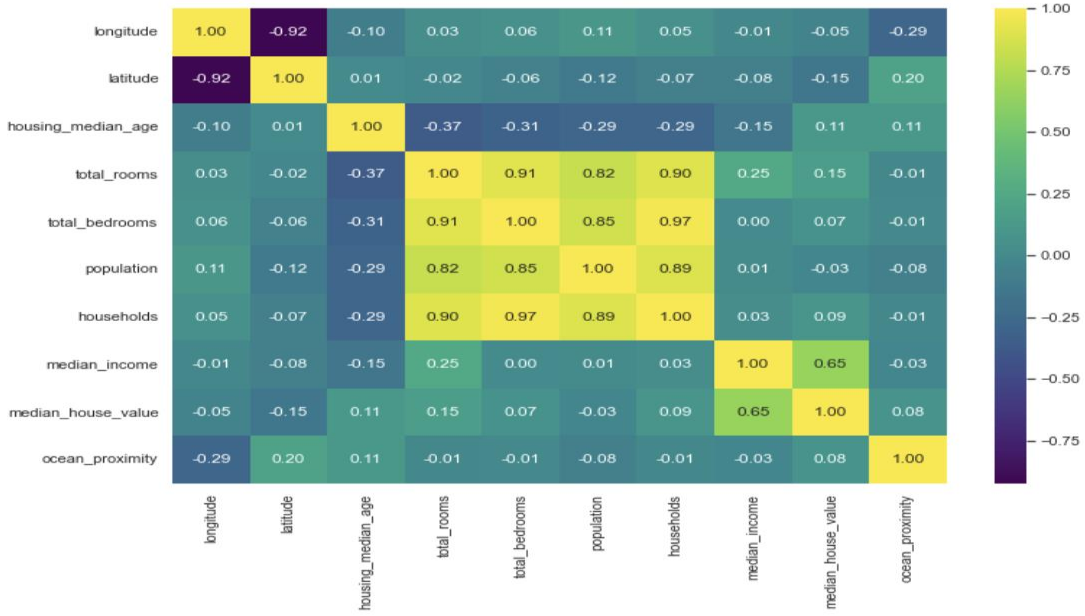
```
# total_bedrooms
Distribution2(columne='total_bedrooms',data=df,i=0)
df[df['total_bedrooms']>=3000].shape
df=df[df['total_bedrooms']<3000]
Distribution2(columne='total_bedrooms',data=df,i=1)
```



Further can also be used to analyzed for other data if required

DROP UNIMPORTANT COLUMNS

```
plt.figure(figsize=(12, 8))
sns.heatmap(data.corr(),annot=True,fmt='.2f',cmap='viridis')
plt.show()
```



Unnecessary data is dropped

FINDINGS

Some of the conclusions that we get from this analysis are:

- It shows the lowest housing pricing in California
- It shows the highest housing pricing in California
- It helps the buyer to buy a property based on their preference by analyzing the data

SUGGESTIONS / RECOMMENDATIONS

The success of any real estate agencies depends on its availability of the property. Here are few suggestions which is purely based on subjective & objective data analysis of the Housing prices.

The availability of a particular property to a potential buyer can be increased if agent provides with a proper marketing strategy. Can provide better view of room and other related data

The planning of needed number of available properties can improve customer interaction. This can be accomplished by using the predictive study of the data and hiring real estate agents from the local agencies.

CONCLUSION

The objective of this analysis was to extract useful information for the housing prices based on the California housing prices data.

This study analyzed the Housing prices currently. The analysis review suggested the highest and lowest prices. At the same time, it also suggests the best pricing for rooms and other in the particular area. The relation between the two provided a better evaluation in the pricing.

The study used a sample data from the whole dataset of Housing prices. The implication of the study is that prices and customer patterns can be used to better increase business efficiencies and profits.

LIMITATIONS

In every research undertaken there are certain unavoidable limitations. This research too has the same. This includes the fact that the data is of a certain year, the current trends of data might be not accessible as it is difficult to acquire required information unless the company decides to share it so.

Another major limitation this analysis is that the data used is specific to a month of the year, and trends changes over the year.

BILBLIOGRAPHY

- https://www.kaggle.com/datasets/camnugent/california-housing-prices
- https://www.simplilearn.com/tutorials/data-analytics-tutorial/what-is-data-analytics
- https://matplotlib.org/stable/index.html
- https://pandas.pydata.org/
- https://seaborn.pydata.org/
- https://www.geeksforgeeks.org/