WELCOME TO THE

NAAN MUDHALVAN PROJECT

TEAM ID: NM2023TMID19767

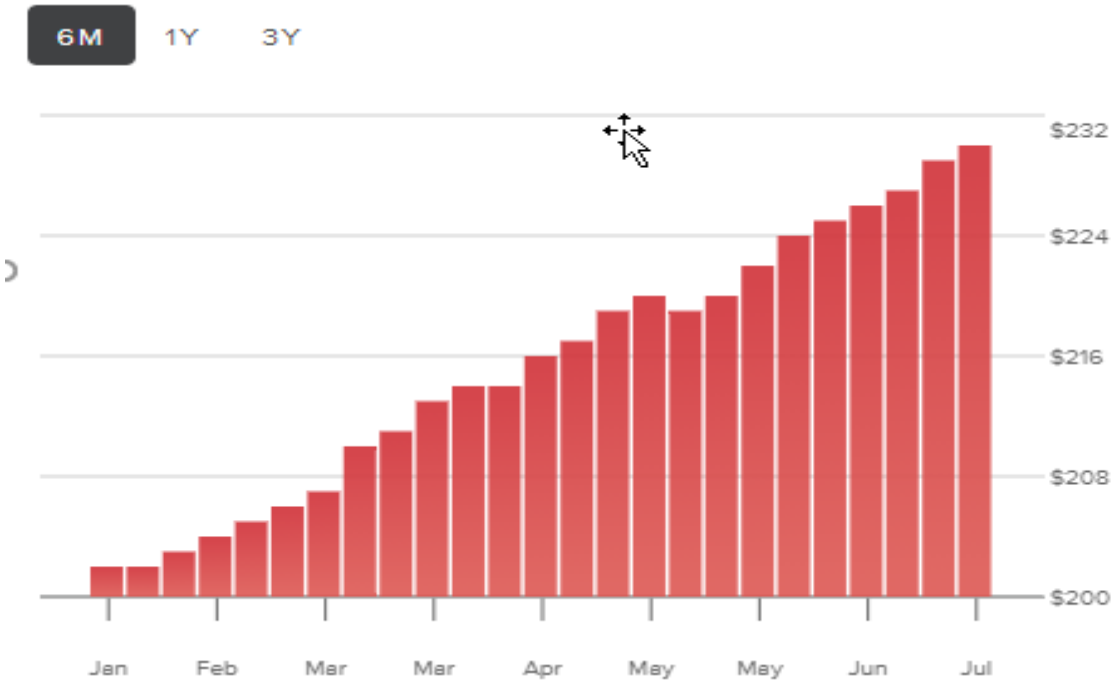
Name: hariharan b

Regno :922520106048

PROGRAM FOR THE HOUSE PRICE PREDICTION

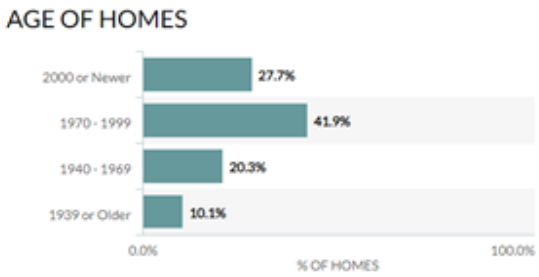
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

# Load the dataset
data = pd.read_csv('housing.csv')

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test =
train_test_split(data.drop('Price', axis=1), data['Price'],
test_size=0.2, random_state=42)

# Train a linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Evaluate the model on the testing set
score = model.score(X_test, y_test)
print(f'R^2 score on testing set: {score:.2f}')
```

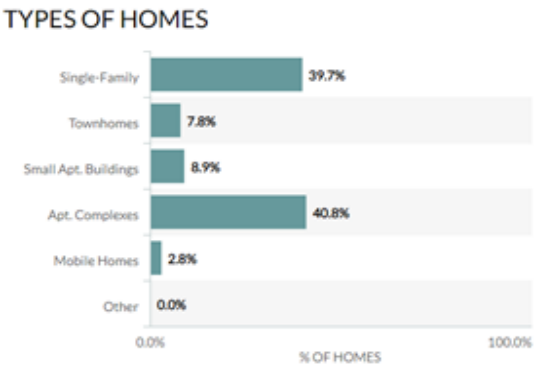
```
# Make a prediction on a new house
new_house = pd.DataFrame({
    'Bedrooms': [3],
    'Bathrooms': [2],
    'Square footage': [1800],
    'Neighborhood': ['Northwest']
})
price = model.predict(new_house)
print(f'Predicted price: ${price[0]:,.2f}')
```

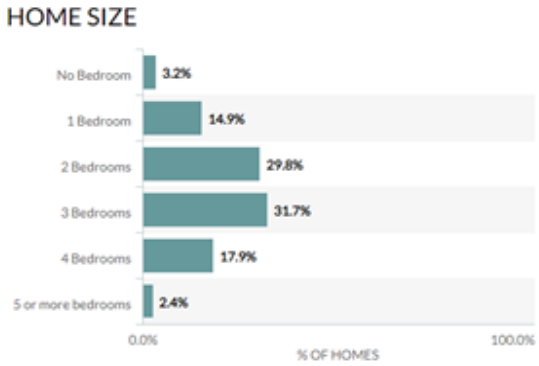


As of July 2021, at the time of this writing, the average price per sqft in Ames is $230.

Exploratory Data Analysis







the home ownership rate is below 35% only.

Dataset looks as follows-

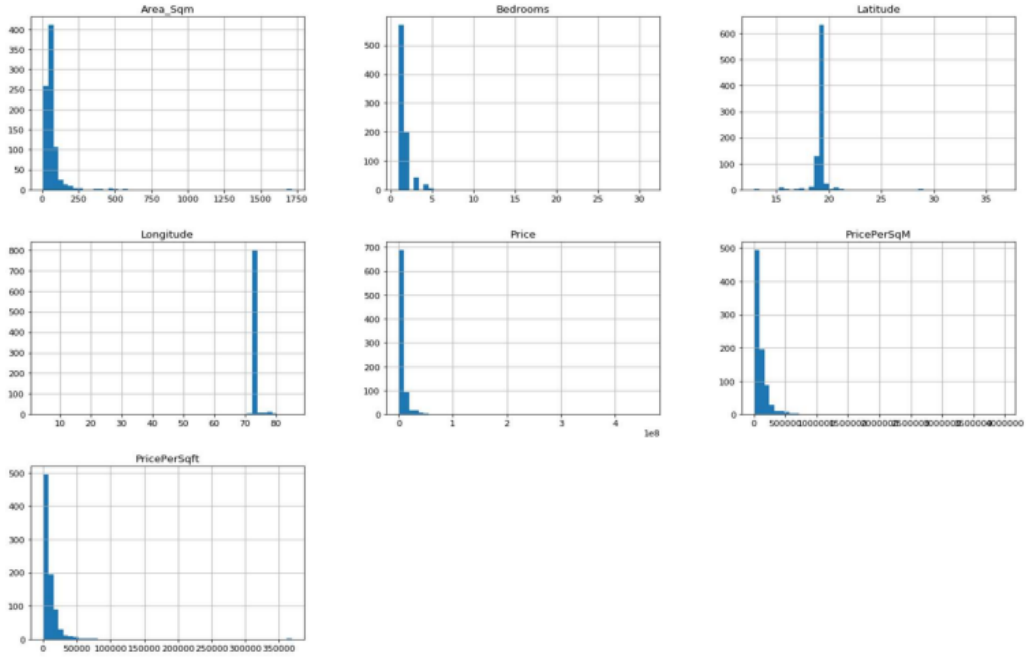
	Price	PricePerSqft	Area_Sqm	Location	Bedrooms	Latitude	Longitude	PricePerSqM
0	13300000	16625	74.32	Kandivali (East)	2	19.210200	72.864891	178885.00
1	9000000	15666	55.74	Ramgad Nagar	1	19.167700	72.949300	168566.16
2	9000000	19148	43.66	Mahakali Caves	1	19.130609	72.873816	206032.48
3	9000000	10588	78.97	Louis Wadi	2	19.126005	72.825052	113926.88
4	100000000	20000	464.51	Barrister Nath Pai Nagar	5	19.075014	72.907571	215200.00

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 840 entries, 0 to 839
Data columns (total 6 columns):
Price          840 non-null int64
Area_Sqm       840 non-null float64
Bedrooms       840 non-null int64
Latitude       840 non-null float64
Longitude      840 non-null float64
PricePerSqM    840 non-null float64
dtypes: float64(4), int64(2)
memory usage: 39.5 KB
```

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working…..

Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyse massive amounts of information and make data-driven decisions



Data Selection

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity. The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.

	Price	Area_Sqm	Bedrooms	Latitude	Longitude	PricePerSqM
0	13300000	74.32	2	19.210200	72.864891	178885.00
1	9000000	55.74	1	19.167700	72.949300	168566.16
2	9000000	43.66	1	19.130609	72.873816	206032.48
3	9000000	78.97	2	19.126005	72.825052	113926.88
4	100000000	464.51	5	19.075014	72.907571	215200.00

Correlation Heatmap