

Assignment

March 11, 2024

1 TensorGo Assignment

author: Anupam Ashish Minz

dataset: California Housing Price

dataset_url: <https://www.kaggle.com/datasets/camnugent/california-housing-prices/data?select=housing.csv>

```
[7]: from langchain_community.llms import Ollama
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
```

```
[8]: df = pd.read_csv("housing.csv")
```

```
[9]: 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
```

```
[10]: df.dropna(inplace=True)
```

```
[11]: df.ocean_proximity.unique()
```

```
[11]: array(['NEAR BAY', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND'],  
        dtype=object)
```

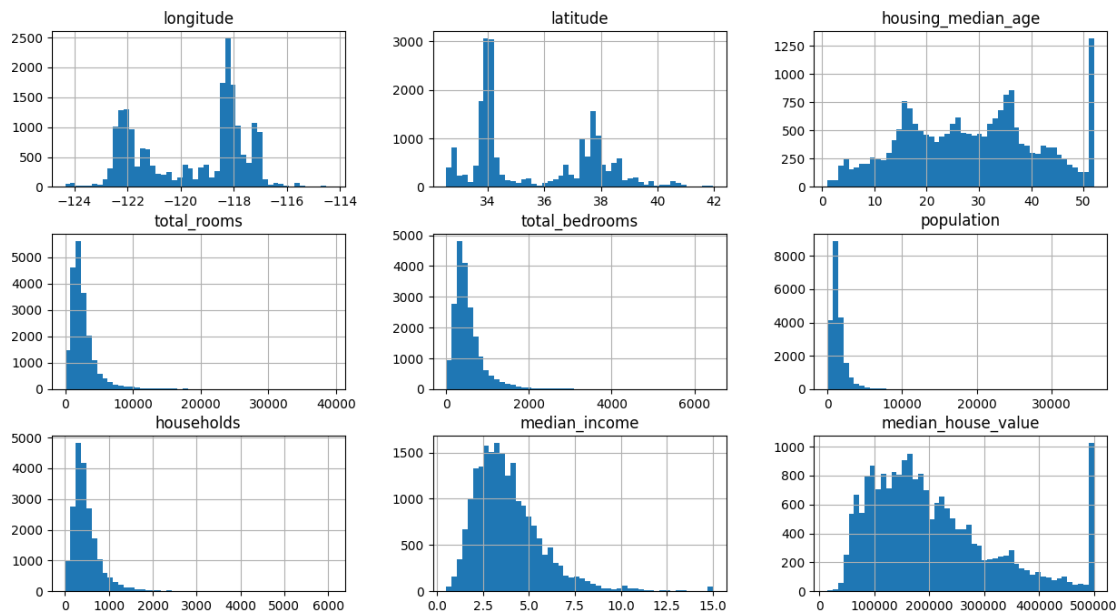
1.1 graphs

1.1.1 Histogram

The following graphs are histograms, they are used here to show the distribution of the data

```
[12]: df.hist(figsize=(15, 8), bins=50)
```

```
[12]: array([[<Axes: title={'center': 'longitude'}>,  
          <Axes: title={'center': 'latitude'}>,  
          <Axes: title={'center': 'housing_median_age'}>],  
        [<Axes: title={'center': 'total_rooms'}>,  
          <Axes: title={'center': 'total_bedrooms'}>,  
          <Axes: title={'center': 'population'}>],  
        [<Axes: title={'center': 'households'}>,  
          <Axes: title={'center': 'median_income'}>,  
          <Axes: title={'center': 'median_house_value'}>]], dtype=object)
```

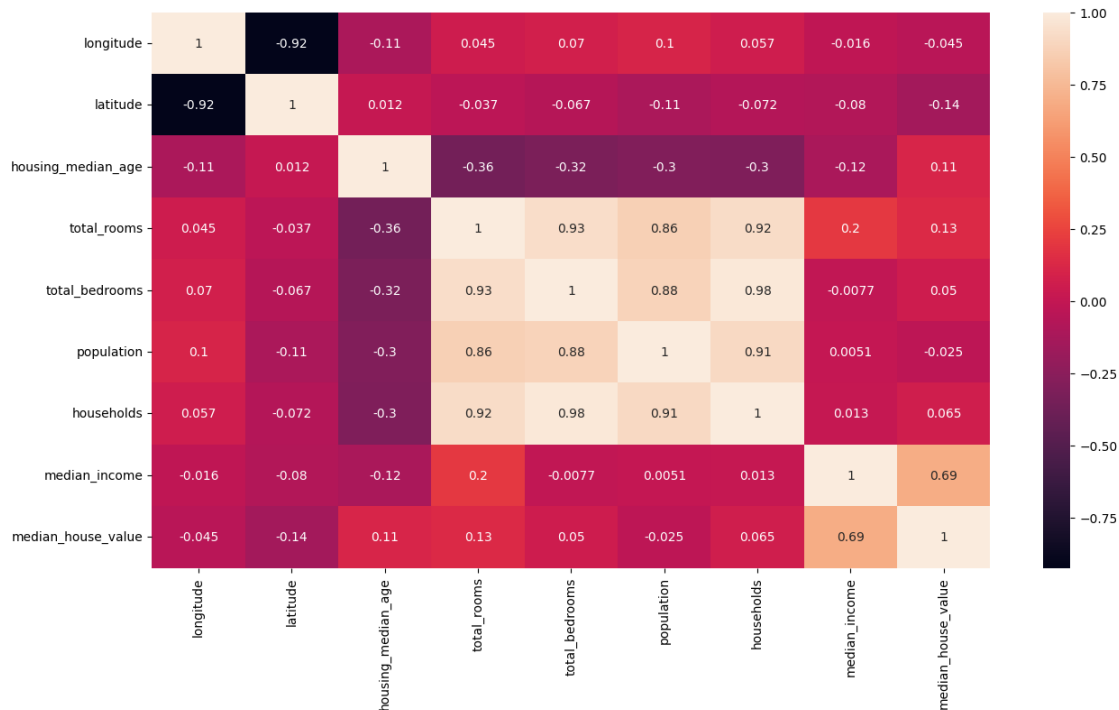


1.1.2 Heatmap

The following is a heatmap, it is used here to show the correlation of attributes with respect to each other

```
[13]: ndf = df.select_dtypes(include=[np.number])
plt.figure(figsize=(15, 8))
sns.heatmap(ndf.corr(), annot=True)
```

[13]: <Axes: >



1.2 llm

As median_house_value is the only value that is of important to us, we are only generating the correlations with respect to that partical value

```
[14]: corr = ndf.drop(['median_house_value'], axis=1).corrwith(df.median_house_value)
print(corr)
```

```
longitude      -0.045398
latitude       -0.144638
housing_median_age  0.106432
total_rooms     0.133294
total_bedrooms  0.049686
population     -0.025300
households      0.064894
median_income   0.688355
dtype: float64
```

We are doing the same with the ocean proximity values, as the values are strings we have to use special functions

```
[16]: ocorr = df.ocean_proximity.str.get_dummies().corrwith(df.median_house_value)
print(ocorr)
```

```
<1H OCEAN    0.257614
INLAND      -0.484787
ISLAND       0.023525
NEAR BAY     0.160526
NEAR OCEAN   0.140378
dtype: float64
```

```
[21]: description = df.describe()
print(description)
```

	longitude	latitude	housing_median_age	total_rooms \
count	20433.000000	20433.000000	20433.000000	20433.000000
mean	-119.570689	35.633221	28.633094	2636.504233
std	2.003578	2.136348	12.591805	2185.269567
min	-124.350000	32.540000	1.000000	2.000000
25%	-121.800000	33.930000	18.000000	1450.000000
50%	-118.490000	34.260000	29.000000	2127.000000
75%	-118.010000	37.720000	37.000000	3143.000000
max	-114.310000	41.950000	52.000000	39320.000000

	total_bedrooms	population	households	median_income \
count	20433.000000	20433.000000	20433.000000	20433.000000
mean	537.870553	1424.946949	499.433465	3.871162
std	421.385070	1133.208490	382.299226	1.899291
min	1.000000	3.000000	1.000000	0.499900
25%	296.000000	787.000000	280.000000	2.563700
50%	435.000000	1166.000000	409.000000	3.536500
75%	647.000000	1722.000000	604.000000	4.744000
max	6445.000000	35682.000000	6082.000000	15.000100

	median_house_value
count	20433.000000
mean	206864.413155
std	115435.667099
min	14999.000000
25%	119500.000000
50%	179700.000000
75%	264700.000000
max	500001.000000

Here we are injecting the description of the data and the correlation values calculated above into the llm

```
[17]: system_message = f"""
you are currently working on the calaifornia housing price dataset
some key indication in the dataset are as follows
```

```
{description}
the correlation of median house values with respect to give paramters are as_
↳follows
{corr}
and the correlation of median house values with respect to proximity of ocean is_
↳as follows
{ocorr}
"""
```

```
[18]: llm = Ollama(model="mistral", system=system_message)
for s in llm.stream("what are some characteristics of the give dataset"):
    print(s, end="")
```

Based on the provided information, here are some characteristics of the California housing price dataset:

1. The dataset contains information about 20,433 houses or housing units in California.
2. The longitude and latitude coordinates provide the geographical location of each house.
3. Housing-related features include median age (in years), total rooms, total bedrooms, and household size.
4. Demographic features include population and median income for the neighborhood or area where each house is located.
5. The dataset also includes a categorical variable "proximity of ocean," which indicates whether a house is located near the ocean (INLAND, NEAR BAY, NEAR OCEAN, <1H OCEAN, or ISLAND).
6. The primary outcome variable is the median housing value or price for each house.
7. The mean median housing age is 28.6 years with a standard deviation of 12.59 years.
8. The mean total rooms and total bedrooms are 2,636.5 and 537.9, respectively.
9. The mean population size for the neighborhood or area is 1,424.9 persons with a standard deviation of 1,133.2.
10. The mean median income for the area is \$3,871.
11. The correlation between median housing values and most features (longitude, latitude, housing_median_age, total_rooms, population, and households) are weak to moderate. However, median housing values have a strong positive correlation with median income and proximity to the ocean (INLAND, NEAR BAY, NEAR OCEAN, <1H OCEAN, or ISLAND).
12. The median housing price is \$206,864.41, but it ranges from a minimum of \$14,999 to a maximum of \$500,001.

```
[19]: for s in llm.stream("which condition affect the median house price the most"):
    print(s, end="")
```

Based on the correlation values provided, it appears that the proximity of a housing unit to the ocean or a body of water (as represented by the variables

"<1H OCEAN," INLAND," ISLAND," NEAR BAY," and NEAR OCEAN) has a stronger relationship with median house prices than other features such as longitude, latitude, housing age, total rooms, total bedrooms, population, households, or median income. Specifically, houses near the ocean or bodies of water tend to have higher median house prices (as indicated by a positive correlation), while houses inland tend to have lower median house prices (as indicated by a negative correlation). However, it's important to keep in mind that correlation does not imply causation, and there may be other factors at play that are influencing housing prices in these areas. It would be useful to explore additional data and perform further analysis to better understand the relationship between housing prices and these variables.

[]: