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

[1]:
```
from langchain_community.llms import Ollama
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
```

[2]:
```
df = pd.read_csv("housing.csv")
```

[3]:
```
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
```

[4]:
```
df.dropna(inplace=True)
```

[5]:
```
df.ocean_proximity.unique()
```

[5]:
```
array(['NEAR BAY', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND'],
      dtype=object)
```

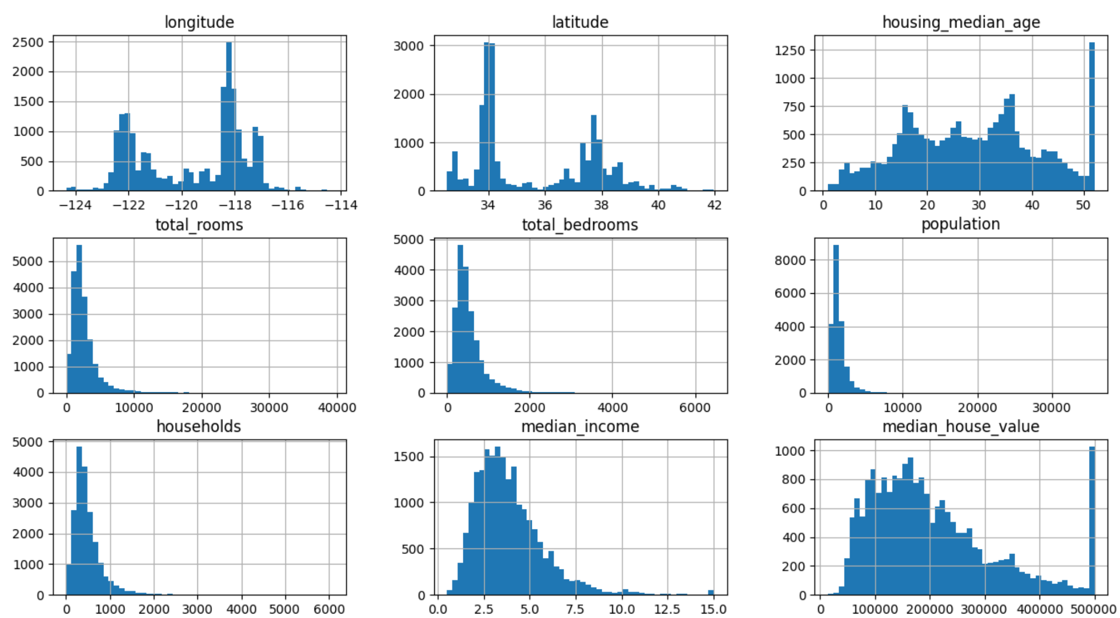
1.1 graphs

1.1.1 Histogram

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

[6]:
```
df.hist(figsize=(15, 8), bins=50)
```

[6]:
```
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)
```

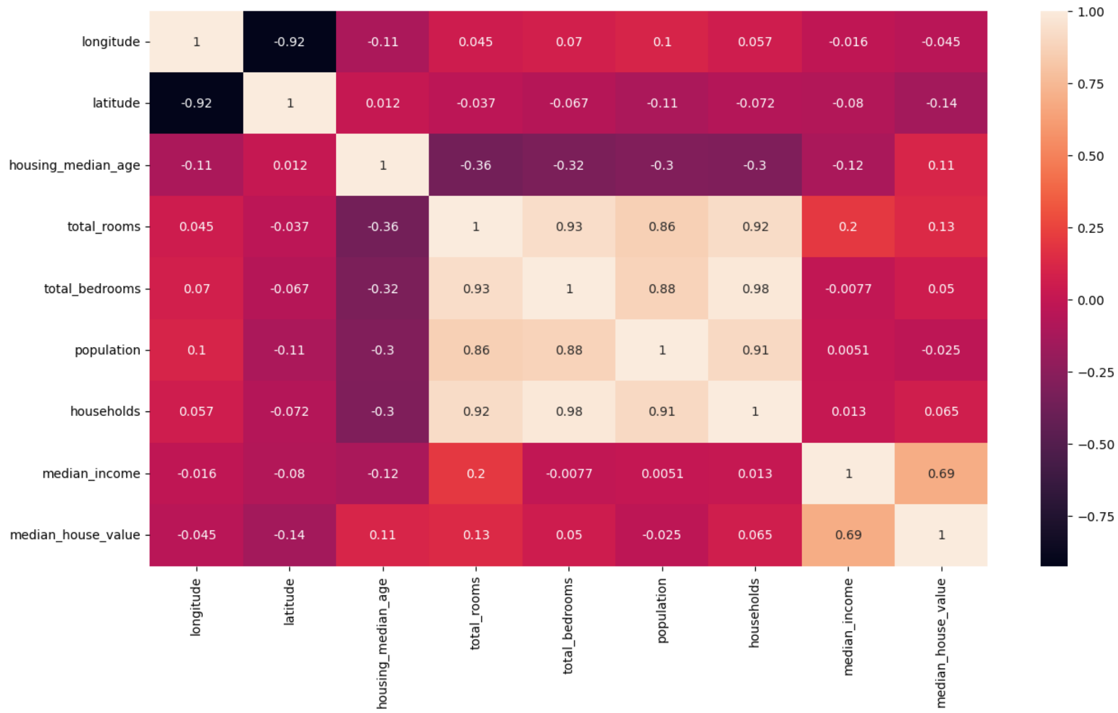


1.1.2 Heatmap

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

[7]:
```python
ndf = df.select_dtypes(include=[np.number])
plt.figure(figsize=(15, 8))
sns.heatmap(ndf.corr(), annot=True)
```

[7]:
```
<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 particual value

[14]:
```python
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

[15]:
```python
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
```

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

[17]:
```python
system_message = f"""
you are currently working on the calafornia housing price dataset
some key indication in the dataset are as follows
{df.describe()}
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}
"""
```

[11]:
```python
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 housing units in California.
Each row represents a single housing unit.

2. The dataset includes several features such as longitude, latitude, housing
median age, total rooms, total bedrooms, population, households, median income,
median house value, and proximity to the ocean (1H OCEAN, INLAND, ISLAND, NEAR
BAY, and NEAR OCEAN).

3. The mean values of longitude and latitude are -119.57 and 35.63,
respectively. The standard deviations for longitude and latitude are 2.00 and
2.14, respectively. This indicates that the housing units in the dataset are
spread out across a large geographical area.

4. The mean values of housing median age, total rooms, and total bedrooms are
28.63, 2636.5, and 537.9, respectively. The standard deviations for housing
median age, total rooms, and total bedrooms are 12.6, 2185.3, and 421.4,
respectively. This suggests that there is a significant variation in the number
```

```
of rooms and bedrooms across different housing units.

5. The mean value of median income is 3.87, indicating that the average
household income is relatively low.

6. The correlation analysis shows that the proximity to the ocean (1H OCEAN) has
a positive correlation with median house values, while being inland has a
negative correlation. This suggests that houses located near the ocean tend to
have higher median house values than those located inland.

7. The correlation between median house values and other features such as
housing median age, total rooms, total bedrooms, population, households, and
median income is relatively weak. However, there is a moderate positive
correlation between median house values and median income. This suggests that
higher income households tend to live in houses with higher median values.
```

[12]:

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

```
 Based on the provided correlation coefficients, it appears that the proximity
of a housing unit to the ocean or a body of water (as indicated by the "NEAR
OCEAN" and "NEAR BAY" features) has a positive correlation with median house
prices. This means that houses that are closer to bodies of water tend to have
higher median house prices compared to those that are further inland.

The correlation coefficient between median house values and proximity to the
ocean is 0.140378, which is relatively strong compared to some of the other
features in the dataset. Additionally, the correlation coefficient for "INLAND"
is negative (-0.484787), indicating that houses located inland tend to have
lower median house prices compared to those near bodies of water or the ocean.

However, it's important to note that correlation does not necessarily imply
causation. Other factors such as location within a city or county, neighborhood
quality, access to amenities, and demographic characteristics can also
significantly impact housing prices. Therefore, while proximity to water appears
to have an effect on median house prices in this dataset, it's just one of many
potential contributing factors.
```

[ ]:

```

```