ce-prediction-project-priyal-desai

March 26, 2024

1 California House Price Prediction Variable Description

1.1 Overview of Dataset

This is the dataset used in the second chapter of Aurélien Géron's recent book 'Hands-On Machine learning with Scikit-Learn and TensorFlow'. It serves as an excellent introduction to implementing machine learning algorithms because it requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size between being to toyish and too cumbersome.

The data contains information from the 1990 California census. So although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introductory dataset for teaching people about the basics of machine learning.

1.2 Description of Variables

1.3 Longitude : Longitude lines run vertically on maps measuring the distance east or west

1.4 Latitude : Latitude lines run horizontally on maps and globes, measuring the distance North to South of the equator.

1.5 While referring to the location of the Pacific Ocean along California's coast, it would be refering to Longitude.

1. longitude: A measure of how far west a house is; a higher value is farther west
2. latitude: A measure of how far north a house is; a higher value is farther north
3. housingMedianAge: Median age of a house within a block; a lower number is a newer building
4. totalRooms: Total number of rooms within a block
5. totalBedrooms: Total number of bedrooms within a block
6. population: Total number of people residing within a block
7. households: Total number of households, a group of people residing within a home unit, for a block
8. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
9. medianHouseValue: Median house value for households within a block (measured in US Dollars)

10. oceanProximity: Location of the house w.r.t ocean/sea

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
```

[2]:
```
import pandas as pd

file_path = "C:\\Users\\priya\\Documents\\House_Price_Prediction_California.
 ↪xlsx"

data = pd.read_excel(file_path)
print(data)
```

```
       longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
0        -122.23     37.88                  41          880           129.0
1        -122.22     37.86                  21         7099          1106.0
2        -122.24     37.85                  52         1467           190.0
3        -122.25     37.85                  52         1274           235.0
4        -122.25     37.85                  52         1627           280.0
…            …        …                    …            …              …
20635    -121.09     39.48                  25         1665           374.0
20636    -121.21     39.49                  18          697           150.0
20637    -121.22     39.43                  17         2254           485.0
20638    -121.32     39.43                  18         1860           409.0
20639    -121.24     39.37                  16         2785           616.0

       population  households  median_income  median_house_value  \
0             322         126         8.3252              452600
1            2401        1138         8.3014              358500
2             496         177         7.2574              352100
3             558         219         5.6431              341300
4             565         259         3.8462              342200
…             …           …              …                   …
20635         845         330         1.5603               78100
20636         356         114         2.5568               77100
20637        1007         433         1.7000               92300
20638         741         349         1.8672               84700
20639        1387         530         2.3886               89400

      ocean_proximity
0            NEAR BAY
1            NEAR BAY
2            NEAR BAY
3            NEAR BAY
4            NEAR BAY
```

```
…                 …
20635           INLAND
20636           INLAND
20637           INLAND
20638           INLAND
20639           INLAND

[20640 rows x 10 columns]
```

[3]:
```
data.head()
```

[3]:
```
   longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
0    -122.23     37.88                  41          880           129.0
1    -122.22     37.86                  21         7099          1106.0
2    -122.24     37.85                  52         1467           190.0
3    -122.25     37.85                  52         1274           235.0
4    -122.25     37.85                  52         1627           280.0

   population  households  median_income  median_house_value ocean_proximity
0         322         126         8.3252              452600        NEAR BAY
1        2401        1138         8.3014              358500        NEAR BAY
2         496         177         7.2574              352100        NEAR BAY
3         558         219         5.6431              341300        NEAR BAY
4         565         259         3.8462              342200        NEAR BAY
```

[4]:
```
data.columns
```

[4]:
```
Index(['longitude', 'latitude', 'housing_median_age', 'total_rooms',
       'total_bedrooms', 'population', 'households', 'median_income',
       'median_house_value', 'ocean_proximity'],
      dtype='object')
```

[5]:
```
data.shape
```

[5]:
```
(20640, 10)
```

[6]:
```
data.describe()
```

[6]:
```
          longitude      latitude  housing_median_age   total_rooms  \
count  20640.000000  20640.000000        20640.000000  20640.000000
mean    -119.569704     35.631861           28.639486   2635.763081
std        2.003532      2.135952           12.585558   2181.615252
min     -124.350000     32.540000            1.000000      2.000000
25%     -121.800000     33.930000           18.000000   1447.750000
50%     -118.490000     34.260000           29.000000   2127.000000
75%     -118.010000     37.710000           37.000000   3148.000000
max     -114.310000     41.950000           52.000000  39320.000000
```

```
       total_bedrooms    population    households  median_income  \
count    20433.000000  20640.000000  20640.000000   20640.000000
mean       537.870553   1425.476744    499.539680       3.870671
std        421.385070   1132.462122    382.329753       1.899822
min          1.000000      3.000000      1.000000       0.499900
25%        296.000000    787.000000    280.000000       2.563400
50%        435.000000   1166.000000    409.000000       3.534800
75%        647.000000   1725.000000    605.000000       4.743250
max       6445.000000  35682.000000   6082.000000      15.000100

       median_house_value
count        20640.000000
mean        206855.816909
std         115395.615874
min          14999.000000
25%         119600.000000
50%         179700.000000
75%         264725.000000
max         500001.000000
```

[7]:
```
data.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  int64
 3   total_rooms         20640 non-null  int64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  int64
 6   households          20640 non-null  int64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  int64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(4), int64(5), object(1)
memory usage: 1.6+ MB
```

[8]:
```
data.isnull().sum()
```

[8]:
```
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
```

```
total_bedrooms        207
population              0
households              0
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

[9]:
```
data.dropna(inplace=True)
```

[10]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 20433 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20433 non-null  float64
 1   latitude            20433 non-null  float64
 2   housing_median_age  20433 non-null  int64
 3   total_rooms         20433 non-null  int64
 4   total_bedrooms      20433 non-null  float64
 5   population          20433 non-null  int64
 6   households          20433 non-null  int64
 7   median_income       20433 non-null  float64
 8   median_house_value  20433 non-null  int64
 9   ocean_proximity     20433 non-null  object
dtypes: float64(4), int64(5), object(1)
memory usage: 1.7+ MB
```

[11]:
```
data.duplicated().sum()
```

[11]:
```
0
```

[12]:
```
data.isnull().sum()
```

[12]:
```
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms        0
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

[13]:
```
from sklearn.model_selection import train_test_split

x = data.drop(['median_house_value'], axis = 1 )
y = data['median_house_value']
```

[14]:
```
x = data.drop(['median_house_value'],axis = 1)
y = data['median_house_value']
```

[15]:
```
x
```

[15]:
```
       longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
0        -122.23     37.88                  41          880           129.0
1        -122.22     37.86                  21         7099          1106.0
2        -122.24     37.85                  52         1467           190.0
3        -122.25     37.85                  52         1274           235.0
4        -122.25     37.85                  52         1627           280.0
…            …        …                    …            …              …
20635    -121.09     39.48                  25         1665           374.0
20636    -121.21     39.49                  18          697           150.0
20637    -121.22     39.43                  17         2254           485.0
20638    -121.32     39.43                  18         1860           409.0
20639    -121.24     39.37                  16         2785           616.0

       population  households  median_income ocean_proximity
0             322         126         8.3252        NEAR BAY
1            2401        1138         8.3014        NEAR BAY
2             496         177         7.2574        NEAR BAY
3             558         219         5.6431        NEAR BAY
4             565         259         3.8462        NEAR BAY
…             …           …              …               …
20635         845         330         1.5603          INLAND
20636         356         114         2.5568          INLAND
20637        1007         433         1.7000          INLAND
20638         741         349         1.8672          INLAND
20639        1387         530         2.3886          INLAND

[20433 rows x 9 columns]
```

[16]:
```
y
```

[16]:
```
0        452600
1        358500
2        352100
3        341300
4        342200
          …
20635     78100
```

```
20636     77100
20637     92300
20638     84700
20639     89400
Name: median_house_value, Length: 20433, dtype: int64
```

1.6 Train-Test Split

[17]:
```
x_train, x_test, y_train, y_test = train_test_split(x,y, test_size = 0.2)
```

[18]:
```
train_data = x_train.join(y_train)
```

[19]:
```
train_data
```

[19]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	\
7092	-118.02	33.92	34	1478	251.0	
2324	-119.73	36.83	14	3348	491.0	
1875	-119.98	38.96	25	2443	444.0	
1357	-121.91	38.02	15	2966	558.0	
9229	-120.16	36.96	18	508	104.0	
…	…	…	…	…	…	
11647	-118.03	33.81	26	3635	567.0	
17420	-120.46	34.64	16	686	217.0	
19883	-119.19	36.34	33	2199	403.0	
11130	-117.93	33.85	33	2489	546.0	
18773	-122.29	40.47	20	2858	612.0	

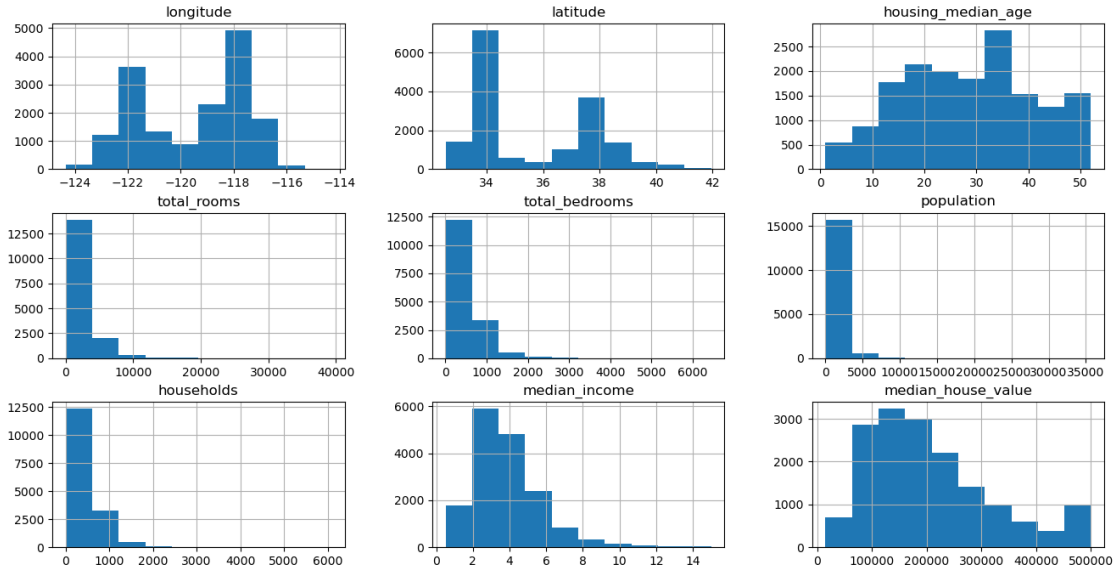
	population	households	median_income	ocean_proximity	\
7092	956	277	5.5238	<1H OCEAN	
2324	1584	493	5.0828	INLAND	
1875	868	342	3.5417	INLAND	
1357	1687	527	3.4817	INLAND	
9229	393	114	3.0000	INLAND	
…	…	…	…	…	
11647	1779	543	5.7089	<1H OCEAN	
17420	614	200	0.8106	NEAR OCEAN	
19883	1245	394	2.7300	INLAND	
11130	1857	444	2.9474	<1H OCEAN	
18773	1422	589	1.9657	INLAND	

	median_house_value
7092	185300
2324	111400
1875	114800
1357	129800
9229	156300
…	…

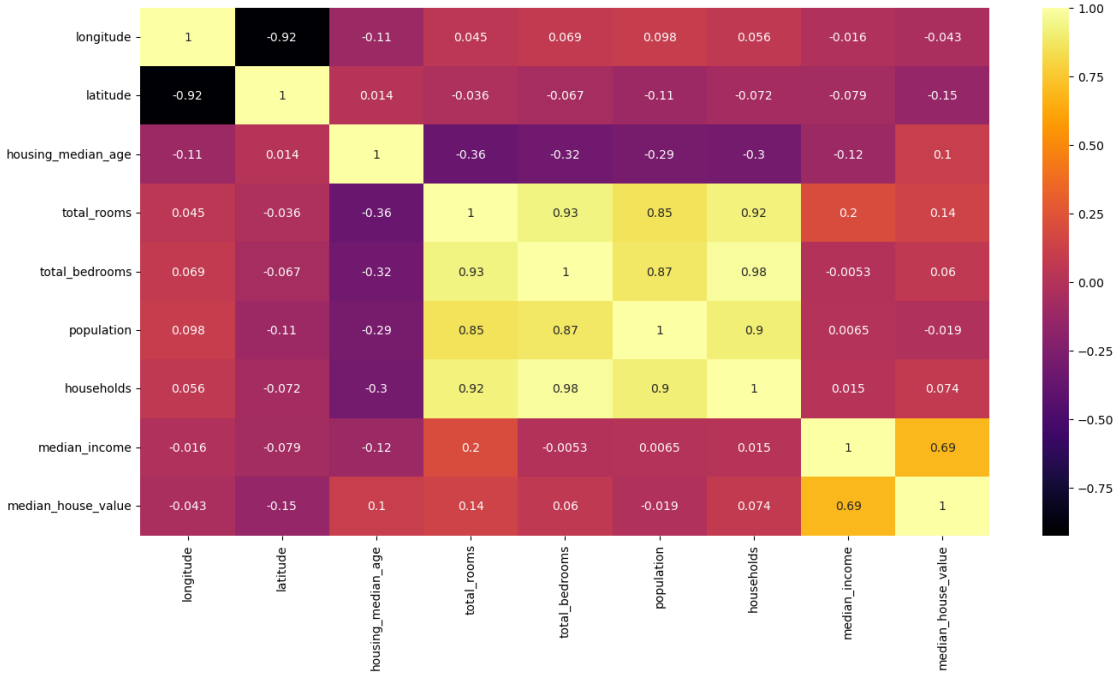
```
11647              237400
17420               83300
19883               96900
11130              178400
18773               63000

[16346 rows x 10 columns]
```

[20]:
```
train_data.hist(figsize = (16,8))
```

[20]:
```
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)
```



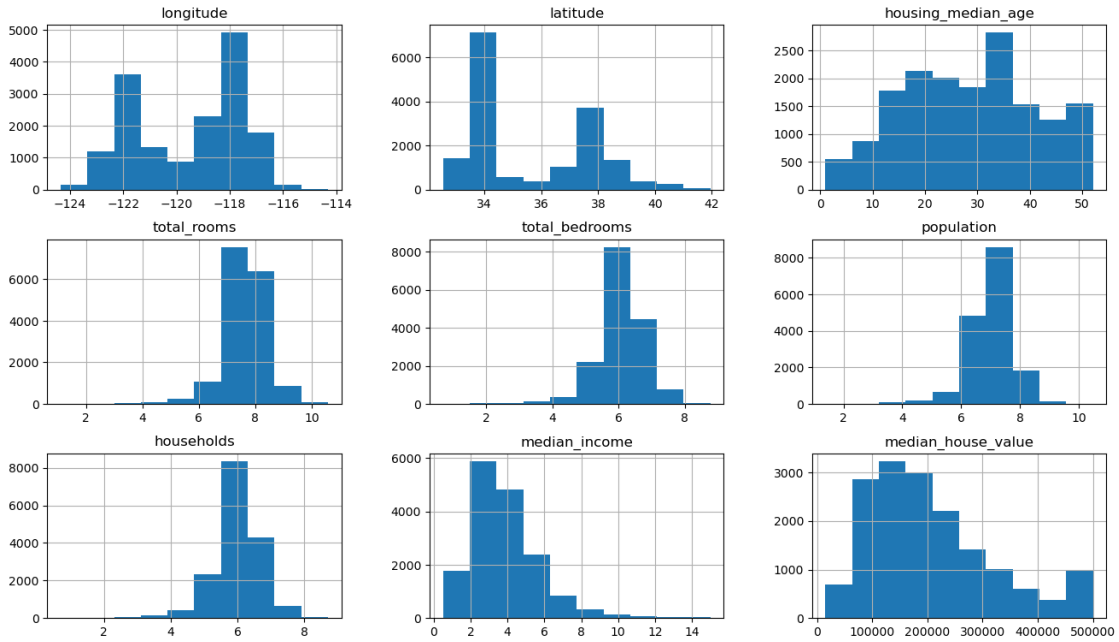
[21]:
```
train_data_no_ocean = train_data.drop('ocean_proximity', axis=1)

plt.figure(figsize=(16, 8))
sns.heatmap(train_data_no_ocean.corr(), annot=True, cmap='inferno')
plt.show()
```



[22]:
```
train_data['total_rooms']=np.log(train_data['total_rooms']+1)
train_data['total_bedrooms']=np.log(train_data['total_bedrooms']+1)
train_data['population']=np.log(train_data['population']+1)
train_data['households']=np.log(train_data['households']+1)
```

[23]:
```
train_data.hist(figsize=(16,9))
```

[23]:
```
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)
```



[24]:
```
train_data.ocean_proximity.value_counts()
```

[24]:
```
ocean_proximity
<1H OCEAN     7257
INLAND        5175
NEAR OCEAN    2092
NEAR BAY      1818
ISLAND           4
Name: count, dtype: int64
```

[25]:
```
pd.get_dummies(train_data.ocean_proximity)
```

[25]:

	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
7092	True	False	False	False	False
2324	False	True	False	False	False
1875	False	True	False	False	False
1357	False	True	False	False	False
9229	False	True	False	False	False
…	…	…	…	…	…
11647	True	False	False	False	False
17420	False	False	False	False	True
19883	False	True	False	False	False
11130	True	False	False	False	False
18773	False	True	False	False	False

[16346 rows x 5 columns]

[26]:
```
dummy_df = pd.get_dummies(train_data.ocean_proximity)
dummy_df = dummy_df.astype(int)

print(dummy_df)
```

```
       <1H OCEAN  INLAND  ISLAND  NEAR BAY  NEAR OCEAN
7092           1       0       0         0           0
2324           0       1       0         0           0
1875           0       1       0         0           0
1357           0       1       0         0           0
9229           0       1       0         0           0
…            …     …     …       …         …
11647          1       0       0         0           0
17420          0       0       0         0           1
19883          0       1       0         0           0
11130          1       0       0         0           0
18773          0       1       0         0           0

[16346 rows x 5 columns]
```

[27]:
```
## Join train data with dummy df of ocean proximity
train_data = train_data.join(dummy_df)
train_data.drop('ocean_proximity', axis=1, inplace=True)
```

[28]:
```
train_data
```

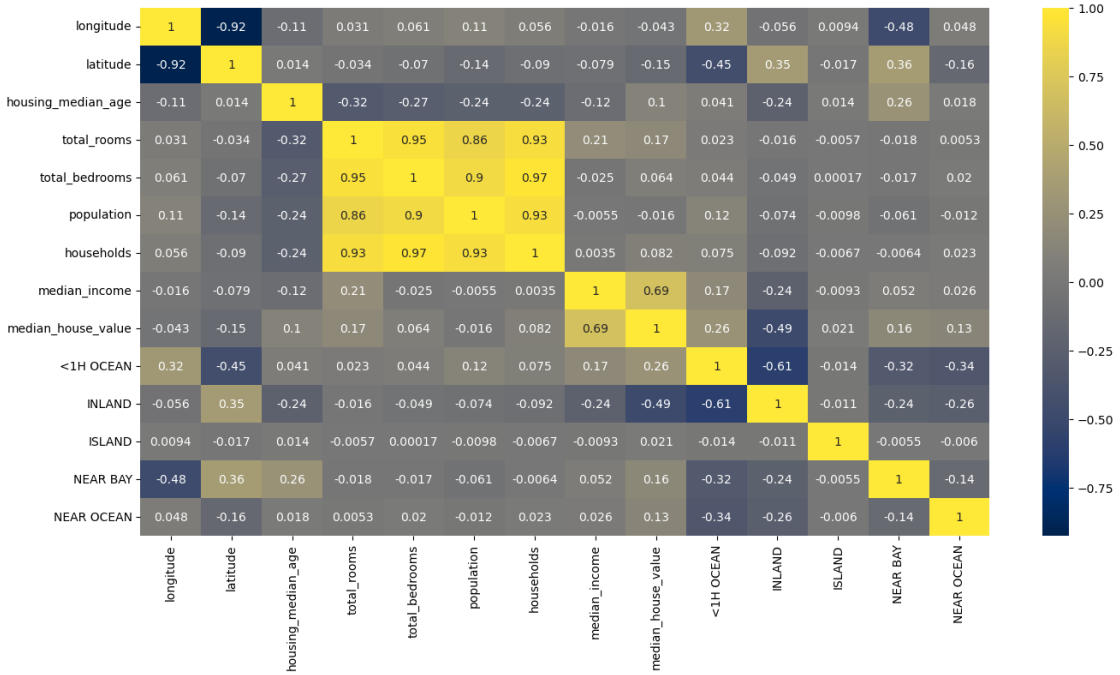
[28]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	\
7092	-118.02	33.92	34	7.299121	5.529429	
2324	-119.73	36.83	14	8.116417	6.198479	
1875	-119.98	38.96	25	7.801391	6.098074	
1357	-121.91	38.02	15	7.995307	6.326149	
9229	-120.16	36.96	18	6.232448	4.653960	
…	…	…	…	…	…	
11647	-118.03	33.81	26	8.198639	6.342121	
17420	-120.46	34.64	16	6.532334	5.384495	
19883	-119.19	36.34	33	7.696213	6.001415	
11130	-117.93	33.85	33	7.820038	6.304449	
18773	-122.29	40.47	20	7.958227	6.418365	

	population	households	median_income	median_house_value	<1H OCEAN	\
7092	6.863803	5.627621	5.5238	185300	1	
2324	7.368340	6.202536	5.0828	111400	0	
1875	6.767343	5.837730	3.5417	114800	0	
1357	7.431300	6.269096	3.4817	129800	0	
9229	5.976351	4.744932	3.0000	156300	0	
…	…	…	…	…	…	
11647	7.484369	6.298949	5.7089	237400	1	

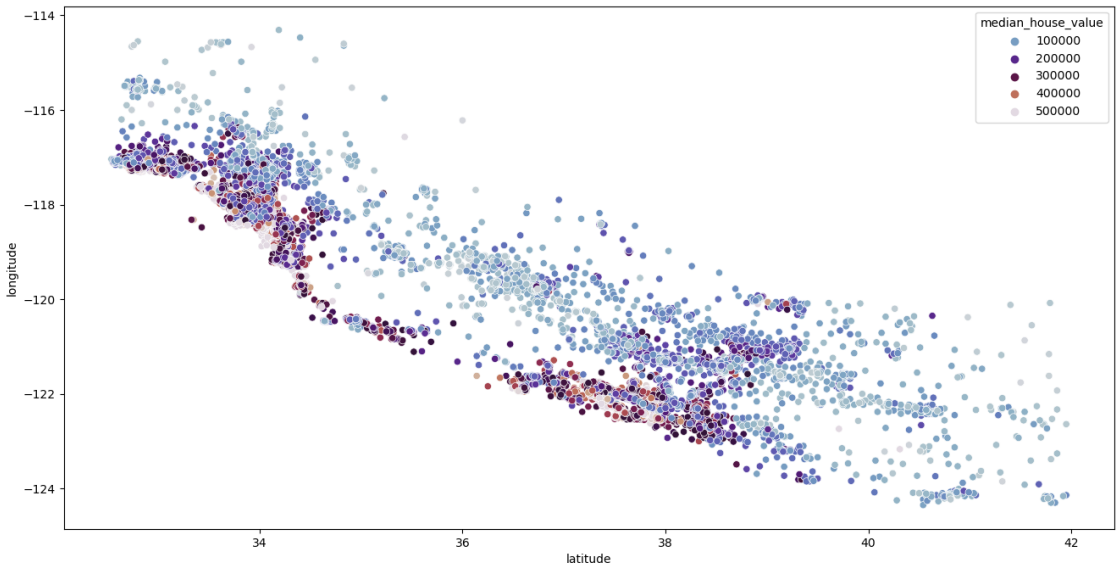
```
17420     6.421622     5.303305          0.8106              83300           0
19883     7.127694     5.978886          2.7300              96900           0
11130     7.527256     6.098074          2.9474             178400           1
18773     7.260523     6.380123          1.9657              63000           0

       INLAND  ISLAND  NEAR BAY  NEAR OCEAN
7092        0       0         0           0
2324        1       0         0           0
1875        1       0         0           0
1357        1       0         0           0
9229        1       0         0           0
...       ...     ...       ...         ...
11647       0       0         0           0
17420       0       0         0           1
19883       1       0         0           0
11130       0       0         0           0
18773       1       0         0           0

[16346 rows x 14 columns]
```

[29]:
```
plt.figure(figsize=(16, 8))
sns.heatmap(train_data.corr(), annot=True, cmap='cividis')
plt.show()
```



[30]:
```
plt.figure(figsize = (16,8))
sns.scatterplot(x = "latitude", y = "longitude", data = train_data, hue =␣
↪"median_house_value", palette = "twilight")
```

[30]: <Axes: xlabel='latitude', ylabel='longitude'>



1.7 In the above Scatterplot it can be observed that the houses those are on the coast in California tend to have higher value than the house which are close to land.

2 *Feature Engineering*

[31]:
```
train_data['bedroom_ratio'] = train_data['total_bedrooms']/
↪train_data['total_rooms']
train_data['household_rooms']= train_data['total_rooms']/
↪train_data['households']
```

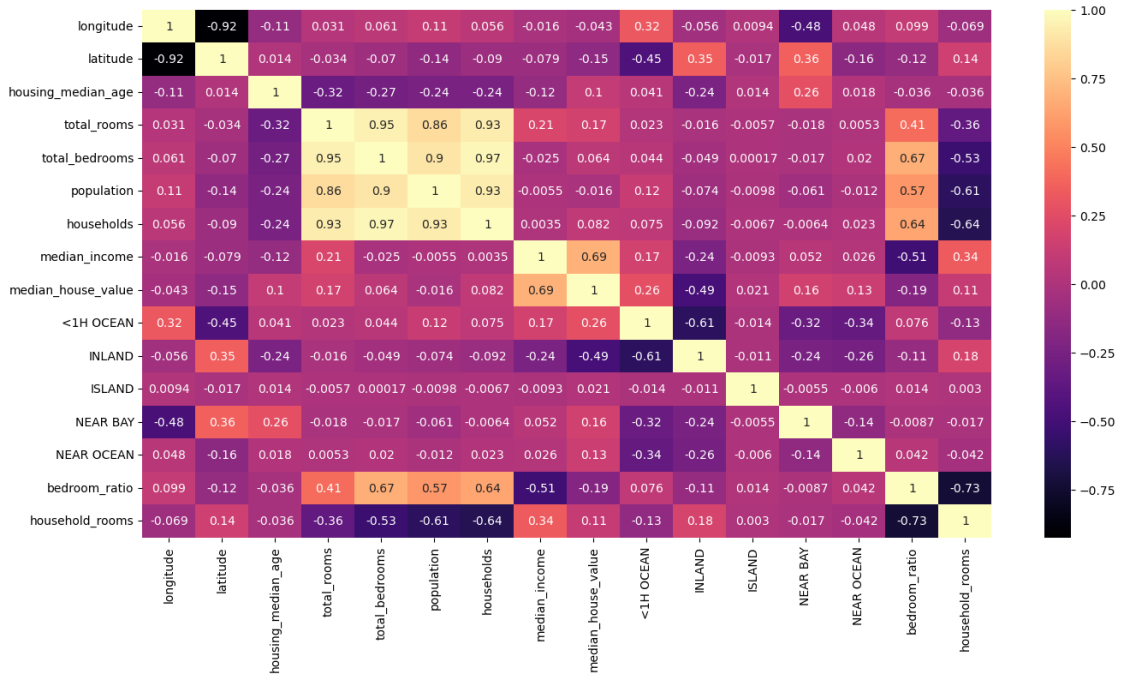
[32]:
```
train_data
```

[32]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	\
7092	-118.02	33.92	34	7.299121	5.529429	
2324	-119.73	36.83	14	8.116417	6.198479	
1875	-119.98	38.96	25	7.801391	6.098074	
1357	-121.91	38.02	15	7.995307	6.326149	
9229	-120.16	36.96	18	6.232448	4.653960	
…	…	…	…	…	…	
11647	-118.03	33.81	26	8.198639	6.342121	

```
17420    -120.46     34.64                  16     6.532334        5.384495
19883    -119.19     36.34                  33     7.696213        6.001415
11130    -117.93     33.85                  33     7.820038        6.304449
18773    -122.29     40.47                  20     7.958227        6.418365

       population  households  median_income  median_house_value  <1H OCEAN  \
7092     6.863803    5.627621         5.5238              185300          1
2324     7.368340    6.202536         5.0828              111400          0
1875     6.767343    5.837730         3.5417              114800          0
1357     7.431300    6.269096         3.4817              129800          0
9229     5.976351    4.744932         3.0000              156300          0
…             …           …              …                   …          …
11647    7.484369    6.298949         5.7089              237400          1
17420    6.421622    5.303305         0.8106               83300          0
19883    7.127694    5.978886         2.7300               96900          0
11130    7.527256    6.098074         2.9474              178400          1
18773    7.260523    6.380123         1.9657               63000          0

       INLAND  ISLAND  NEAR BAY  NEAR OCEAN  bedroom_ratio  household_rooms
7092        0       0         0           0       0.757547         1.297017
2324        1       0         0           0       0.763696         1.308564
1875        1       0         0           0       0.781665         1.336374
1357        1       0         0           0       0.791233         1.275352
9229        1       0         0           0       0.746731         1.313496
…       …       …         …           …       …                …
11647       0       0         0           0       0.773558         1.301588
17420       0       0         0           1       0.824283         1.231748
19883       1       0         0           0       0.779788         1.287232
11130       0       0         0           0       0.806192         1.282378
18773       1       0         0           0       0.806507         1.247347

[16346 rows x 16 columns]
```

[52]:
```
plt.figure(figsize=(16, 8))
sns.heatmap(train_data.corr(), annot=True, cmap='magma')
plt.show()
```



2.1 Implementation of Simple Linear Regression

[34]:
```
from sklearn.linear_model import LinearRegression


x_train, y_train = train_data.drop(['median_house_value'], axis=1),␣
↪train_data['median_house_value']

reg = LinearRegression()

reg.fit(x_train, y_train)
```

[34]:
```
LinearRegression()
```

[35]:
```
train_data
```

[35]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	\
7092	-118.02	33.92	34	7.299121	5.529429	
2324	-119.73	36.83	14	8.116417	6.198479	
1875	-119.98	38.96	25	7.801391	6.098074	
1357	-121.91	38.02	15	7.995307	6.326149	
9229	-120.16	36.96	18	6.232448	4.653960	
…	…	…	…	…	…	
11647	-118.03	33.81	26	8.198639	6.342121	
17420	-120.46	34.64	16	6.532334	5.384495	

```
19883    -119.19     36.34                33     7.696213        6.001415
11130    -117.93     33.85                33     7.820038        6.304449
18773    -122.29     40.47                20     7.958227        6.418365

       population  households  median_income  median_house_value  <1H OCEAN  \
7092     6.863803    5.627621         5.5238              185300          1
2324     7.368340    6.202536         5.0828              111400          0
1875     6.767343    5.837730         3.5417              114800          0
1357     7.431300    6.269096         3.4817              129800          0
9229     5.976351    4.744932         3.0000              156300          0
…             …           …              …                   …          …
11647    7.484369    6.298949         5.7089              237400          1
17420    6.421622    5.303305         0.8106               83300          0
19883    7.127694    5.978886         2.7300               96900          0
11130    7.527256    6.098074         2.9474              178400          1
18773    7.260523    6.380123         1.9657               63000          0

       INLAND  ISLAND  NEAR BAY  NEAR OCEAN  bedroom_ratio  household_rooms
7092        0       0         0           0       0.757547         1.297017
2324        1       0         0           0       0.763696         1.308564
1875        1       0         0           0       0.781665         1.336374
1357        1       0         0           0       0.791233         1.275352
9229        1       0         0           0       0.746731         1.313496
…       …       …         …           …       …              …
11647       0       0         0           0       0.773558         1.301588
17420       0       0         0           1       0.824283         1.231748
19883       1       0         0           0       0.779788         1.287232
11130       0       0         0           0       0.806192         1.282378
18773       1       0         0           0       0.806507         1.247347

[16346 rows x 16 columns]
```

2.2 Test_Data Preparation

[36]:
```
test_data = x_test.join(y_test)

test_data['total_rooms']=np.log(test_data['total_rooms']+1)
test_data['total_bedrooms']=np.log(test_data['total_bedrooms']+1)
test_data['population']=np.log(test_data['population']+1)
test_data['households']=np.log(test_data['households']+1)

dummy_df = pd.get_dummies(test_data.ocean_proximity)
dummy_df = dummy_df.astype(int)

test_data = test_data.join(dummy_df)
test_data.drop('ocean_proximity', axis=1, inplace=True)
```

```
test_data['bedroom_ratio'] = test_data['total_bedrooms']/
 ↪test_data['total_rooms']
test_data['household_rooms']= test_data['total_rooms']/test_data['households']
```

[38]:
```
test_data
```

[38]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	\
14105	-117.10	32.75	15	7.792762	6.652863	
20333	-118.99	34.23	9	9.270400	7.388946	
10640	-117.79	33.69	16	8.028781	5.983936	
20423	-119.00	34.08	17	7.508239	6.084499	
18223	-122.08	37.41	20	7.548029	6.124683	
…	…	…	…	…	…	
14958	-116.95	32.76	13	8.620472	6.754604	
8071	-118.17	33.82	50	8.185350	6.542472	
10604	-117.81	33.67	9	7.798113	5.983936	
19123	-122.65	38.24	24	7.575072	5.739793	
2230	-119.77	36.84	15	7.629976	6.023448	

	population	households	median_income	median_house_value	<1H OCEAN	\
14105	7.659643	6.573680	1.0617	92400	0	
20333	8.482809	7.382124	6.6246	284200	1	
10640	7.151485	5.921578	8.7385	340000	1	
20423	6.361302	5.676754	5.4346	428600	0	
18223	6.975414	6.079933	4.6875	288900	0	
…	…	…	…	…	…	
14958	7.637716	6.603944	4.9528	266200	1	
8071	7.322510	6.480045	5.5106	252200	0	
10604	7.085901	5.955837	7.2025	275000	1	
19123	6.827629	5.749393	4.9500	243600	1	
2230	6.793466	5.937536	3.2569	124400	0	

	INLAND	ISLAND	NEAR BAY	NEAR OCEAN	bedroom_ratio	household_rooms
14105	0	0	0	1	0.853723	1.185449
20333	0	0	0	0	0.797047	1.255790
10640	0	0	0	0	0.745311	1.355852
20423	0	0	0	1	0.810376	1.322629
18223	0	0	1	0	0.811428	1.241466
…	…	…	…	…	…	…
14958	0	0	0	0	0.783554	1.305352
8071	0	0	0	1	0.799290	1.263163
10604	0	0	0	0	0.767357	1.309323
19123	0	0	0	0	0.757721	1.317543
2230	1	0	0	0	0.789445	1.285041

[4087 rows x 16 columns]

[39]:
```
replacement_dict = {True: 1, False: 0}

test_data.replace(replacement_dict, inplace=True)

test_data
```

[39]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	\
14105	-117.10	32.75	15	7.792762	6.652863	
20333	-118.99	34.23	9	9.270400	7.388946	
10640	-117.79	33.69	16	8.028781	5.983936	
20423	-119.00	34.08	17	7.508239	6.084499	
18223	-122.08	37.41	20	7.548029	6.124683	
…	…	…	…	…	…	
14958	-116.95	32.76	13	8.620472	6.754604	
8071	-118.17	33.82	50	8.185350	6.542472	
10604	-117.81	33.67	9	7.798113	5.983936	
19123	-122.65	38.24	24	7.575072	5.739793	
2230	-119.77	36.84	15	7.629976	6.023448	

	population	households	median_income	median_house_value	<1H OCEAN	\
14105	7.659643	6.573680	1.0617	92400	0	
20333	8.482809	7.382124	6.6246	284200	1	
10640	7.151485	5.921578	8.7385	340000	1	
20423	6.361302	5.676754	5.4346	428600	0	
18223	6.975414	6.079933	4.6875	288900	0	
…	…	…	…	…	…	
14958	7.637716	6.603944	4.9528	266200	1	
8071	7.322510	6.480045	5.5106	252200	0	
10604	7.085901	5.955837	7.2025	275000	1	
19123	6.827629	5.749393	4.9500	243600	1	
2230	6.793466	5.937536	3.2569	124400	0	

	INLAND	ISLAND	NEAR BAY	NEAR OCEAN	bedroom_ratio	household_rooms
14105	0	0	0	1	0.853723	1.185449
20333	0	0	0	0	0.797047	1.255790
10640	0	0	0	0	0.745311	1.355852
20423	0	0	0	1	0.810376	1.322629
18223	0	0	1	0	0.811428	1.241466
…	…	…	…	…	…	…
14958	0	0	0	0	0.783554	1.305352
8071	0	0	0	1	0.799290	1.263163
10604	0	0	0	0	0.767357	1.309323
19123	0	0	0	0	0.757721	1.317543
2230	1	0	0	0	0.789445	1.285041

[4087 rows x 16 columns]

[40]:
```
x_test, y_test = test_data.drop(['median_house_value'], axis=1),␣
↪test_data['median_house_value']
```

[41]:
```
reg.score(x_test, y_test)
```

[41]: 0.6538243850997112

[42]:
```
x_test
```

[42]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	\
14105	-117.10	32.75	15	7.792762	6.652863	
20333	-118.99	34.23	9	9.270400	7.388946	
10640	-117.79	33.69	16	8.028781	5.983936	
20423	-119.00	34.08	17	7.508239	6.084499	
18223	-122.08	37.41	20	7.548029	6.124683	
…	…	…	…	…	…	
14958	-116.95	32.76	13	8.620472	6.754604	
8071	-118.17	33.82	50	8.185350	6.542472	
10604	-117.81	33.67	9	7.798113	5.983936	
19123	-122.65	38.24	24	7.575072	5.739793	
2230	-119.77	36.84	15	7.629976	6.023448	

	population	households	median_income	<1H OCEAN	INLAND	ISLAND	\
14105	7.659643	6.573680	1.0617	0	0	0	
20333	8.482809	7.382124	6.6246	1	0	0	
10640	7.151485	5.921578	8.7385	1	0	0	
20423	6.361302	5.676754	5.4346	0	0	0	
18223	6.975414	6.079933	4.6875	0	0	0	
…	…	…	…	…	…	…	
14958	7.637716	6.603944	4.9528	1	0	0	
8071	7.322510	6.480045	5.5106	0	0	0	
10604	7.085901	5.955837	7.2025	1	0	0	
19123	6.827629	5.749393	4.9500	1	0	0	
2230	6.793466	5.937536	3.2569	0	1	0	

	NEAR BAY	NEAR OCEAN	bedroom_ratio	household_rooms
14105	0	1	0.853723	1.185449
20333	0	0	0.797047	1.255790
10640	0	0	0.745311	1.355852
20423	0	1	0.810376	1.322629
18223	1	0	0.811428	1.241466
…	…	…	…	…
14958	0	0	0.783554	1.305352
8071	0	1	0.799290	1.263163
10604	0	0	0.767357	1.309323

```
19123         0           0       0.757721         1.317543
2230          0           0       0.789445         1.285041

[4087 rows x 15 columns]
```

[43]:
```
y_test
```

[43]:
```
14105     92400
20333    284200
10640    340000
20423    428600
18223    288900
          …
14958    266200
8071     252200
10604    275000
19123    243600
2230     124400
Name: median_house_value, Length: 4087, dtype: int64
```

[44]:
```
test_data
```

[44]:
```
       longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
14105    -117.10     32.75                  15     7.792762        6.652863
20333    -118.99     34.23                   9     9.270400        7.388946
10640    -117.79     33.69                  16     8.028781        5.983936
20423    -119.00     34.08                  17     7.508239        6.084499
18223    -122.08     37.41                  20     7.548029        6.124683
…          …       …                     …          …              …
14958    -116.95     32.76                  13     8.620472        6.754604
8071     -118.17     33.82                  50     8.185350        6.542472
10604    -117.81     33.67                   9     7.798113        5.983936
19123    -122.65     38.24                  24     7.575072        5.739793
2230     -119.77     36.84                  15     7.629976        6.023448

       population  households  median_income  median_house_value  <1H OCEAN  \
14105    7.659643    6.573680         1.0617                92400          0
20333    8.482809    7.382124         6.6246               284200          1
10640    7.151485    5.921578         8.7385               340000          1
20423    6.361302    5.676754         5.4346               428600          0
18223    6.975414    6.079933         4.6875               288900          0
…          …          …              …                    …         …
14958    7.637716    6.603944         4.9528               266200          1
8071     7.322510    6.480045         5.5106               252200          0
10604    7.085901    5.955837         7.2025               275000          1
19123    6.827629    5.749393         4.9500               243600          1
2230     6.793466    5.937536         3.2569               124400          0
```

	INLAND	ISLAND	NEAR BAY	NEAR OCEAN	bedroom_ratio	household_rooms
14105	0	0	0	1	0.853723	1.185449
20333	0	0	0	0	0.797047	1.255790
10640	0	0	0	0	0.745311	1.355852
20423	0	0	0	1	0.810376	1.322629
18223	0	0	1	0	0.811428	1.241466
…	…	…	…	…	…	…
14958	0	0	0	0	0.783554	1.305352
8071	0	0	0	1	0.799290	1.263163
10604	0	0	0	0	0.767357	1.309323
19123	0	0	0	0	0.757721	1.317543
2230	1	0	0	0	0.789445	1.285041

[4087 rows x 16 columns]

2.3 Random Forest Model

[45]:
```
from sklearn.ensemble import RandomForestRegressor

forest = RandomForestRegressor()

forest.fit(x_train, y_train)
```

[45]: RandomForestRegressor()

[46]:
```
forest.score(x_test, y_test)
```

[46]: 0.8089425647699034