1901202051.SB.ML1S3.Final

October 15, 2021

0.1 ML

Name: Suryajiraje Bhosale

PRN: 1901202051

School: School of Data Science

Program: B.Sc. Data Science

Year/ Semester: Second Year Semester 3

Subject Name: Machine Learning-I

Subject Code: DS303

Title: Perform an exploratory data analysis on the California Housing dataset. Use appropriate libraries.

Skills/Competencies to be acquired:

1. To gain an understanding of data and find clues from the data.
2. Assess assumptions on which statistical inference will be based.
3. To check the quality of data for further processing and cleaning if necessary.
4. To check for anomalies or outliers that may impact model.
5. Data Visualization.

Duration of activity: 1 Hour

What is the purpose of this activity?

1. Preview data.
2. Check total number of entries and column types.
3. Check any null values.
4. Check duplicate entries.
5. Plot distribution of numeric data (univariate and pairwise joint distribution).
6. Plot count distribution of categorical data.

Steps performed in this activity.

1. EDA
2. Geo-Spatial Analysis
3. Data Visualisation

What resources / materials / equipment / tools did you use for this activity?

1. California Housing Dataset
2. Jupyter Notebook & Relevant Libraries

What skills did you acquire?

1. EDA
2. Data Visualisation

Time taken to complete the activity?

1 Hour

[134]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.image as mpimg

sns.set(color_codes=True)
sns.set_palette(sns.color_palette('muted'))
```

[135]:
```
df = pd.read_csv('/Users/user/Desktop/CalHousing.csv')
```

[136]:
```
df.head()
```

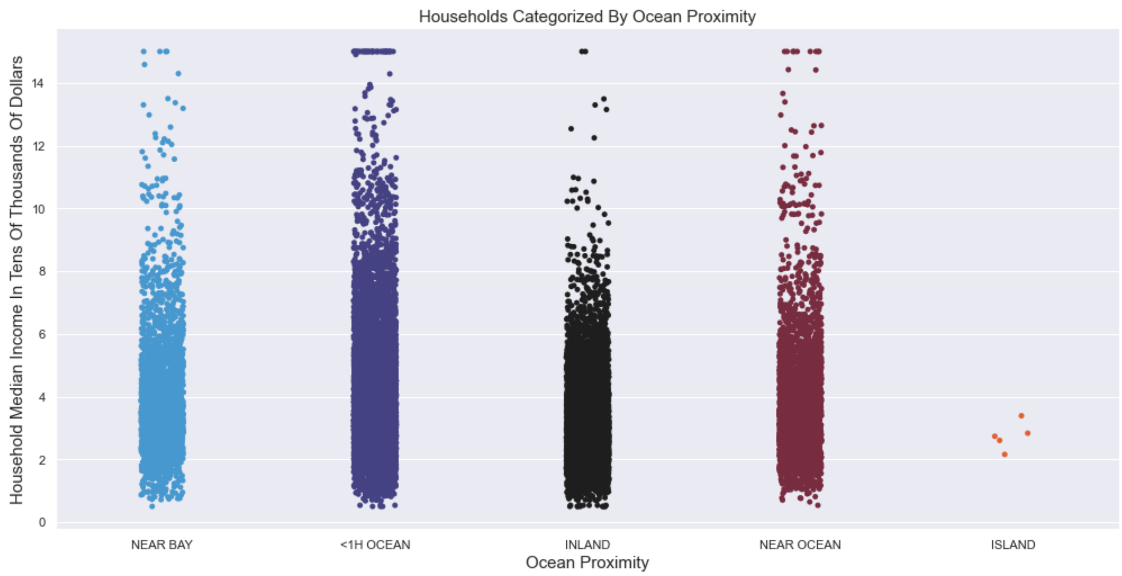
[136]:
```
   longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
0    -122.23     37.88                  41          880           129.0
1    -122.22     37.86                  21         7099          1106.0
2    -122.24     37.85                  52         1467           190.0
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
```

[137]:
```
ocprox = pd.DataFrame(df.ocean_proximity.value_counts())
ocprox
```

[137]:
```
            ocean_proximity
<1H OCEAN              9136
INLAND                 6551
NEAR OCEAN             2658
NEAR BAY               2290
ISLAND                    5
```

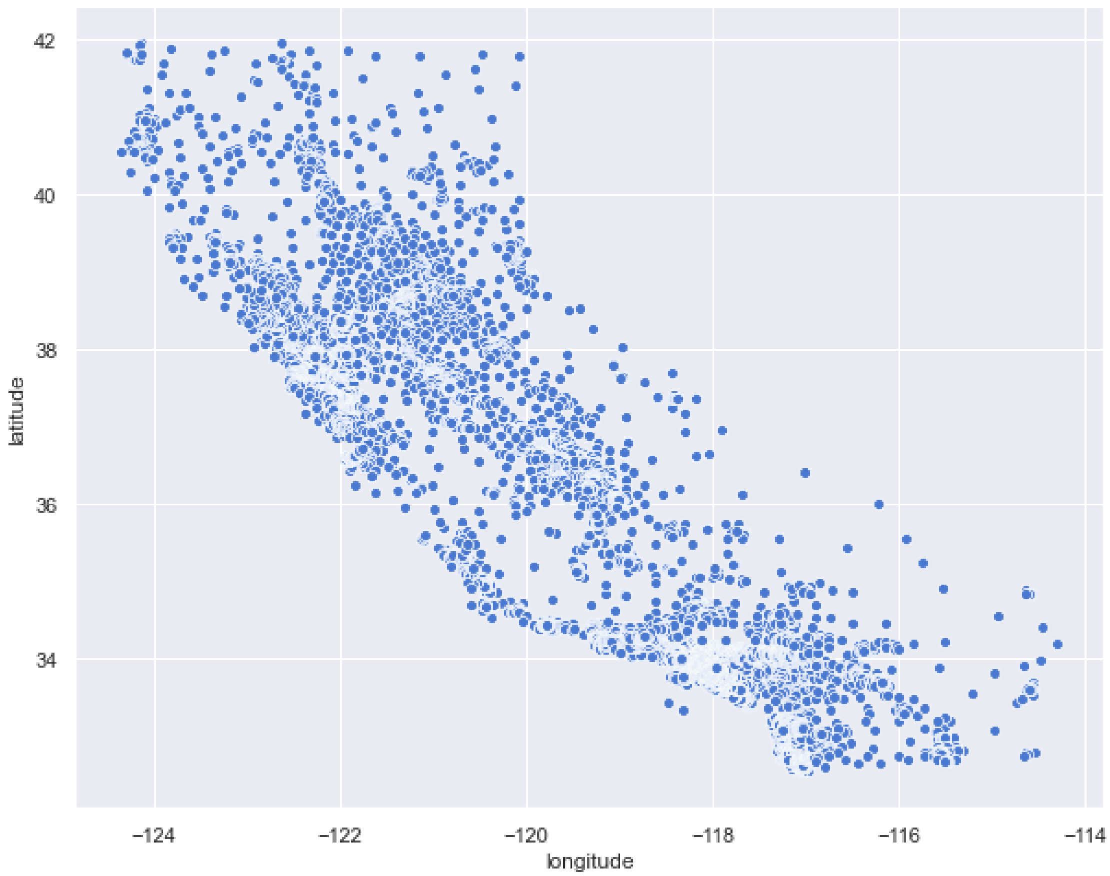
[176]:
```
fig0 = sns.catplot(data = df,x= 'ocean_proximity',y= 'median_income',␣
 ↪palette="icefire",height=7, aspect=2);
plt.title('Households Categorized By Ocean Proximity',fontsize=15 );
plt.ylabel('Household Median Income In Tens Of Thousands Of Dollars',␣
 ↪fontsize=15);
plt.xlabel('Ocean Proximity', fontsize=15);
```



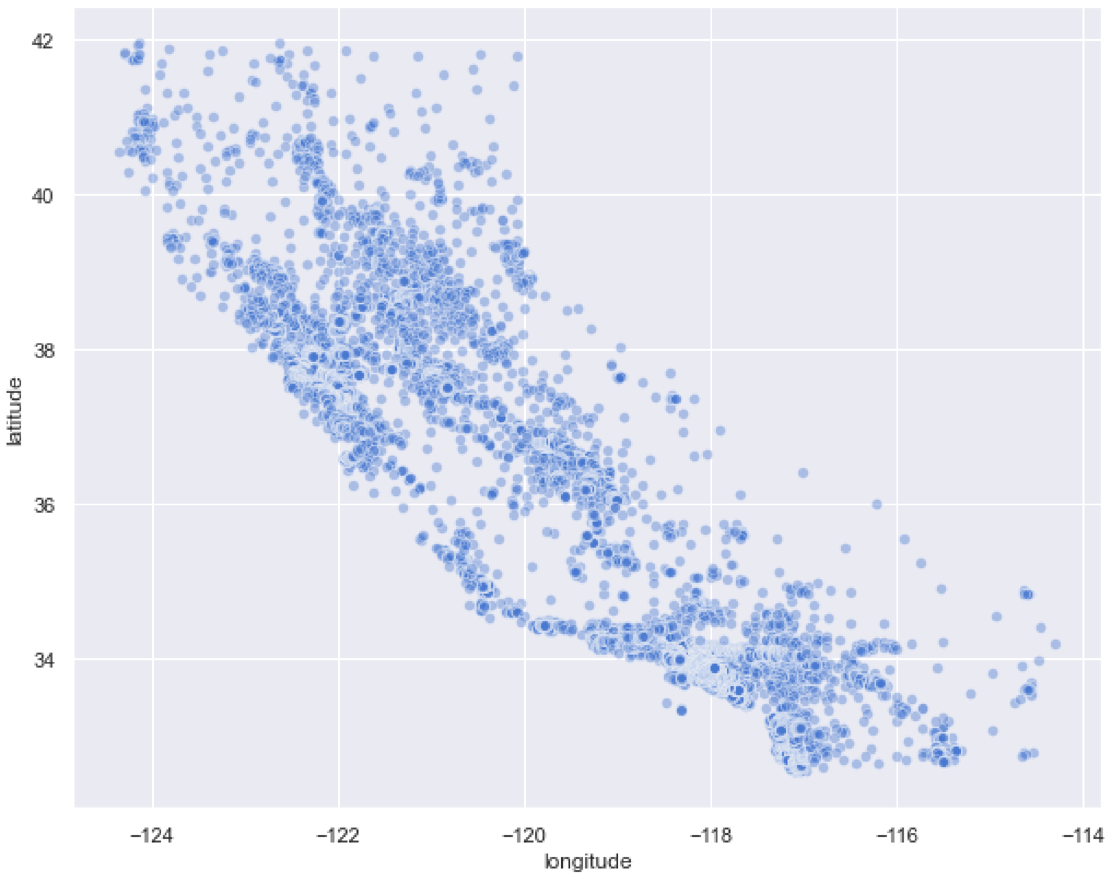
Comment:

The above category plot and the correspoding coprox dataframe depicts that the most inhabited area is the <1H Ocean with 9136 households. From the catplot above we can deduce that the most number of high earning households, setting the floor at $100,000, too live in the area <H Ocean.

[144]:
```
fig_dims = (10, 8)
fig, ax = plt.subplots(figsize=fig_dims)
fig = sns.scatterplot(x=df.longitude, y=df.latitude)
```



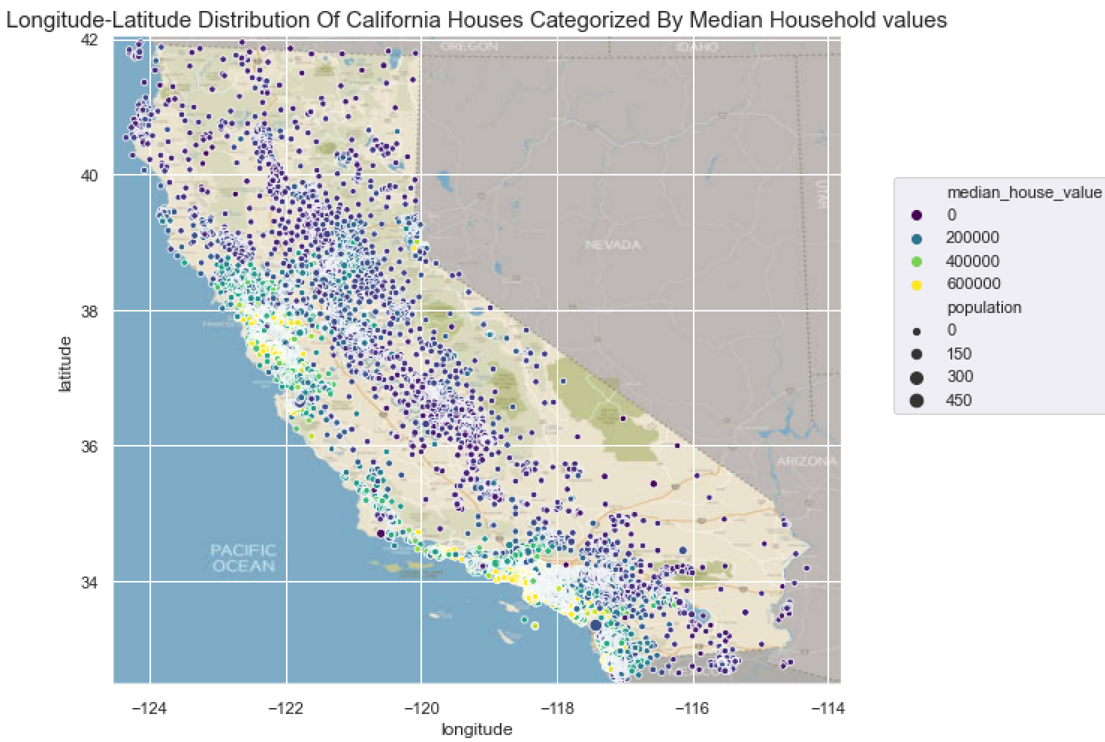
[178]:
```
fig_dims = (10, 8)
fig, ax = plt.subplots(figsize=fig_dims)
fig = sns.scatterplot(x=df.longitude, y=df.latitude, alpha=0.4)
```



Comment:

We have plotted the longitudnal and latitudnal plotting of households on a grid, we can breiefly understand the that strucutre conforms to the shape of california. Another long-lat grid at alpha 0.4 stauration/opacity displays underlying clusters of households.

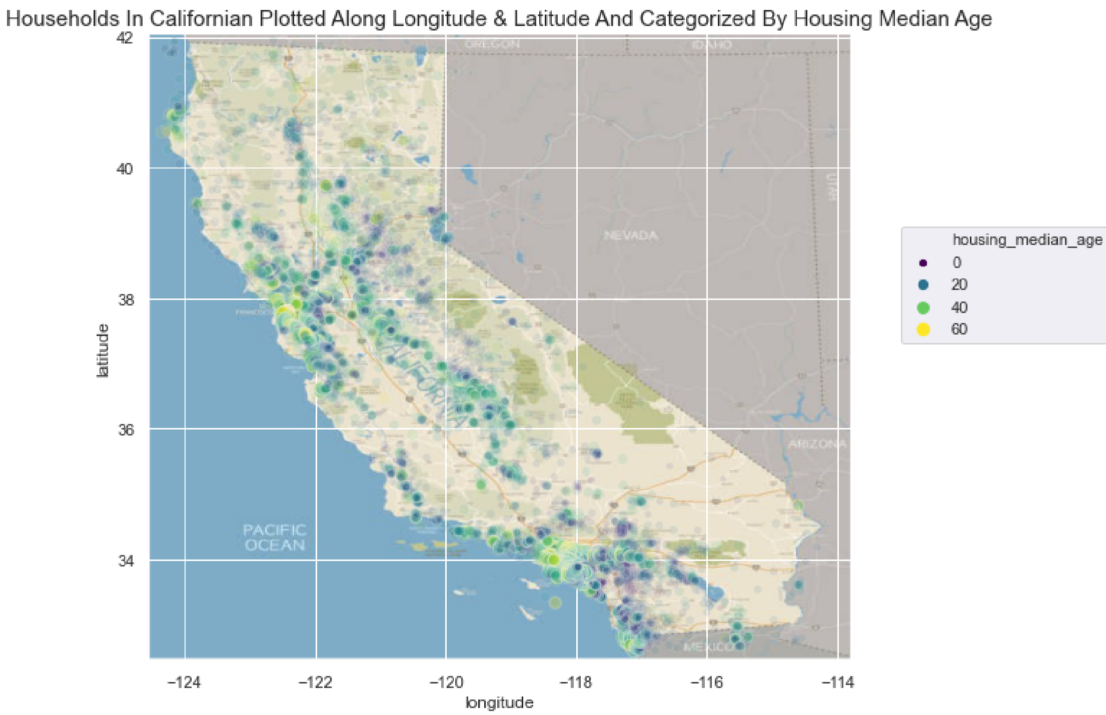
[165]:
```
fig_dims = (10, 8)
fig, ax = plt.subplots(figsize=fig_dims)
fig = sns.scatterplot(x=df.longitude, y=df.latitude, hue=df.median_house_value,␣
↪size=df.population/100,
                      palette='viridis', ax=ax)
fig.legend(loc="center left", bbox_to_anchor=(1.06, 0.6), ncol=1);
california_img=mpimg.imread('/Users/user/Desktop/CaliMap.jpg')
plt.imshow(california_img, extent=[-124.55, -113.80, 32.45, 42.05]);
plt.title('Longitude-Latitude Distribution Of California Houses Categorized By␣
↪Median Household values', fontsize = 15);
```



Comments:

The graphic above shows us the distribution of households hued around median house value and sized about population in population/100. We can see high median housing clustered in anticipated centers of Los Angeles with major clusters of houses valuing near or above $600,000. The cities also show a high population density. Intercity roadways have a dispersed households, where as city suburbs having low median value households with proportionaly high population clusters.

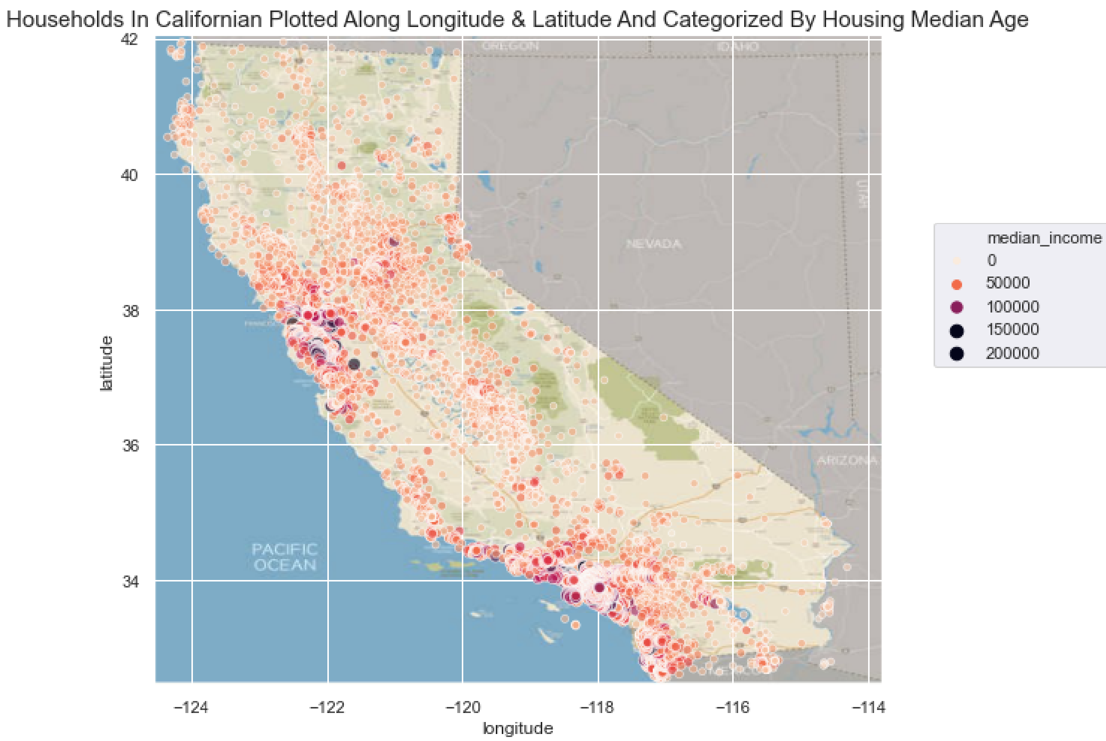
[180]:
```
figdim = (10,8)
fig2, ax = plt.subplots(figsize = figdim)
fig2 = sns.scatterplot(x=df.longitude, y=df.latitude, hue=df.
 ↪housing_median_age,size=df.housing_median_age, palette='viridis', ax=ax,␣
 ↪alpha=0.1)
fig2.legend(loc="center left", bbox_to_anchor=(1.06, 0.6), ncol=1);
plt.imshow(california_img, extent=[-124.55, -113.80, 32.45, 42.05]);
plt.title('Households In Californian Plotted Along Longitude & Latitude And␣
 ↪Categorized By Housing Median Age', fontsize=15);
```



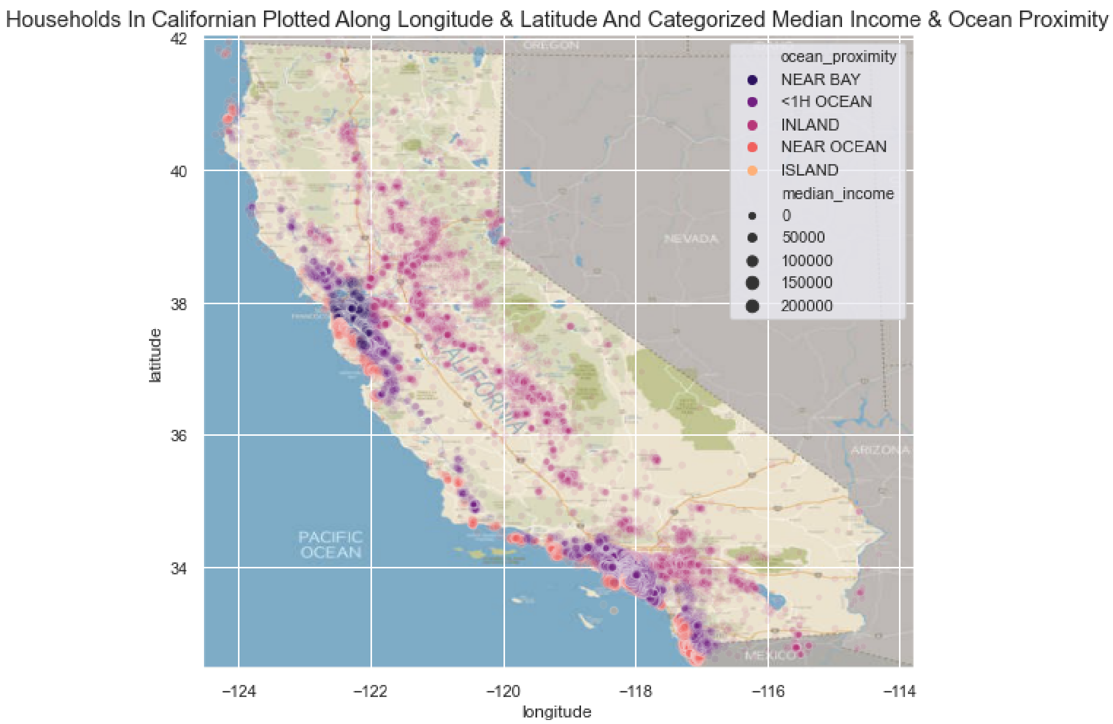
Comments:

In the above graphic we can see that there are heavy clusters of highly aged houses in the cities and suburbs of Los Angeles and San Francisco. The median age of the houses in the inter-metropolis patch, which can be expected. The further away fromt he coastal settlements the households are the more young they appear to be in term of their median age.

[118]:
```
figdim = (10,8)
fig3, ax = plt.subplots(figsize = figdim)
fig3 = sns.scatterplot(x=df.longitude, y=df.latitude, hue=df.
 ↪median_income*10000,size=df.median_income*10000, palette='rocket_r', ax=ax,␣
 ↪alpha=0.6)
fig3.legend(loc="center left", bbox_to_anchor=(1.06, 0.6), ncol=1);
plt.imshow(california_img, extent=[-124.55, -113.80, 32.45, 42.05]);
plt.title('Households In Californian Plotted Along Longitude & Latitude And␣
 ↪Categorized By Housing Median Age', fontsize=15);
```



[181]:
```
fig4, ax = plt.subplots(figsize = figdim)
fig4 = sns.scatterplot(x=df.longitude, y=df.latitude, hue=df.ocean_proximity,␣
↪size=df.median_income*10000, alpha=0.1, ax=ax,
                       palette='magma')
plt.imshow(california_img, extent=[-124.55, -113.80, 32.45, 42.05]);
plt.title('Households In Californian Plotted Along Longitude & Latitude And␣
↪Categorized Median Income & Ocean Proximity', fontsize=15);
```

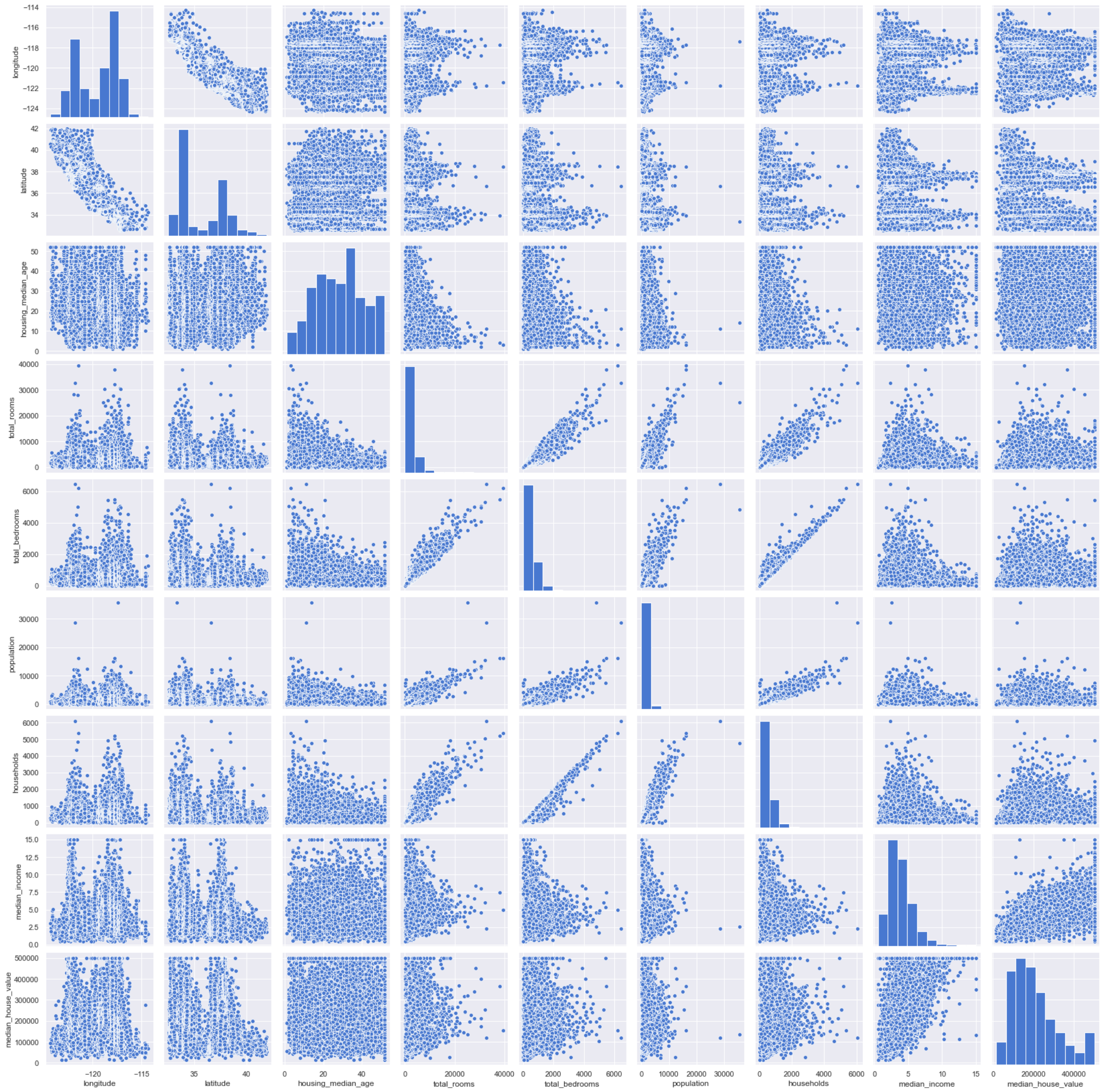


Comment:

The above graphic shows us at alpha 0.1 opacity adjustment the median income of households in the state of California, which has been categorized around the ocean proximity parameter. This aids in visualizing tht high networth households have near ocean and near bay proximity. The big cluster of dark purple on the both north and south ends also depict important geo-economic vitals such as geo-spaital location of income bandwidths of households.

[156]:
```
sns.pairplot(df)
```

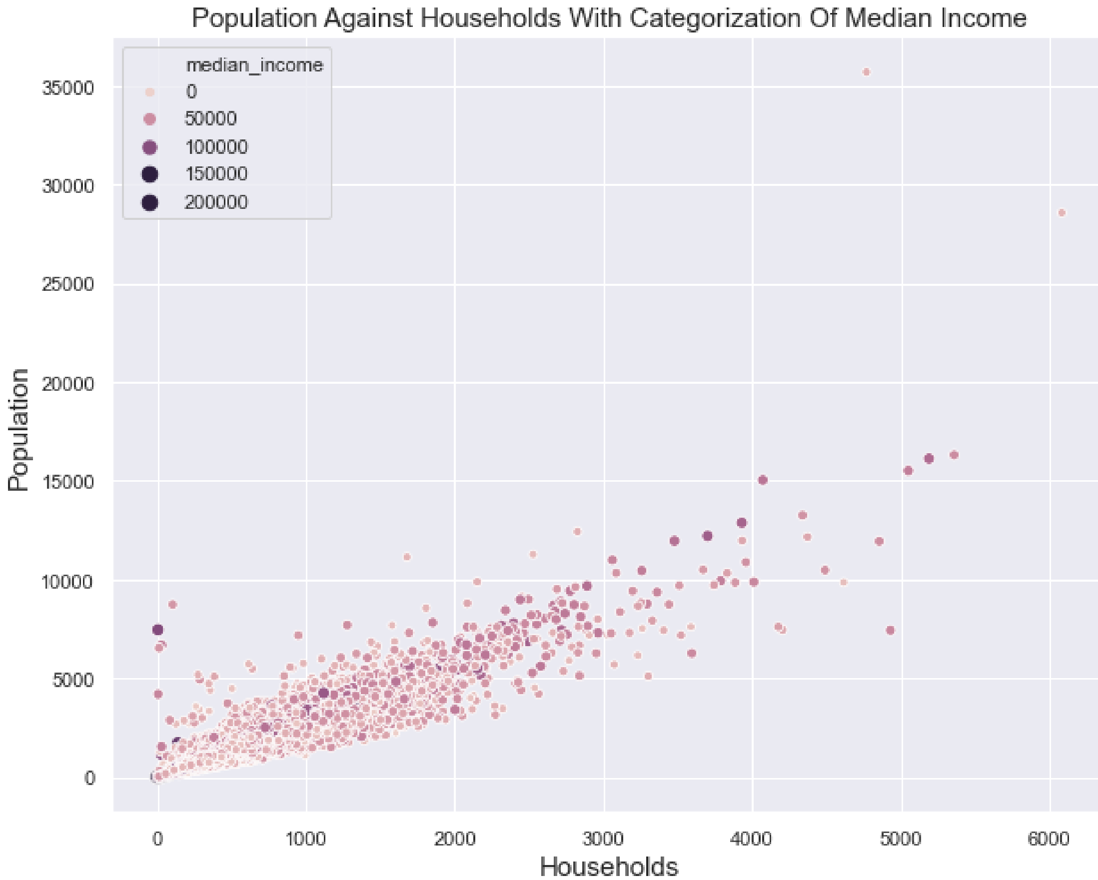
[156]: `<seaborn.axisgrid.PairGrid at 0x14117cc18>`



Comment:

The pairplot above aid aids in visualising the basic relationships between paramenters in the dataset. The ad hoc understanding of the relationships aid in assessing the associations and correlations of the paramters graphically.

[163]:
```
fig_dims = (10, 8)
fig100, ax = plt.subplots(figsize=fig_dims)
fig100 = sns.scatterplot(x=df.households, y=df.population, hue=df.median_income*10000, size=df.median_income*10000)
plt.title('Population Against Households With Categorization Of Median Income', fontsize = 15);
plt.ylabel('Population', fontsize=15);
plt.xlabel('Households', fontsize=15);
```



Comment:

This above graphic shows is the reltionship between household and populations, which is expected to be positively correlated and the categorization of these datapoints based on median income. Many such graphs can be curated using pairplot and categorical parameters can be leveraged to gain a better understand of the distribution of data points.

[ ]:

```

```