Advanced Data Science Training

Final Test: Housing price DEA

By Bobo BUYA

January 2025

Chapter 1: Description

```
In [31]: #Import Data
         data=pd.read_csv('housing.csv')

In [32]: print(data.info())
         print('------------------------------------------------------------------------')
         print(data.columns)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 11 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
 10  AgeClass            20640 non-null  object
dtypes: float64(4), int64(5), object(2)
memory usage: 1.7+ MB
None
------------------------------------------------------------------------
Index(['longitude', 'latitude', 'housing_median_age', 'total_rooms',
       'total_bedrooms', 'population', 'households', 'median_income',
       'median_house_value', 'ocean_proximity', 'AgeClass'],
      dtype='object')
```

```
In [4]: data.nunique()
```

```
Out[4]: longitude             844
        latitude              862
        housing_median_age     52
        total_rooms          5926
        total_bedrooms       1923
        population           3888
        households           1815
        median_income       12928
        median_house_value   3842
        ocean_proximity         5
        dtype: int64
```

The data file came in a CSV format. The data frame generated had 11 Columns and a matrix of shape 20,604 X 11.

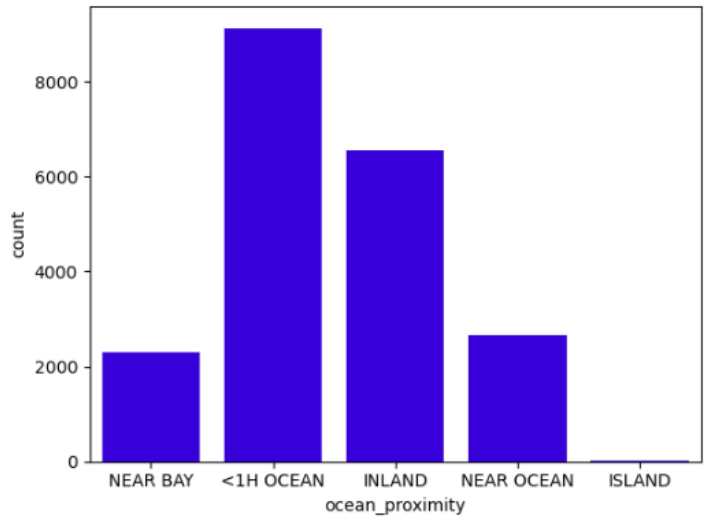
```
In [36]: data.shape
```

```
Out[36]: (20640, 11)
```

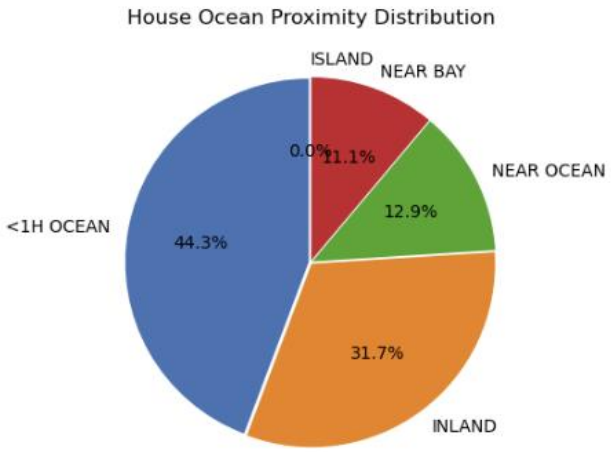
Chapter 2: House Ocean proximity distribution

```
In [9]: sns.countplot(x='ocean_proximity', data=data, color = 'b')
```

```
Out[9]: <AxesSubplot:xlabel='ocean_proximity', ylabel='count'>
```



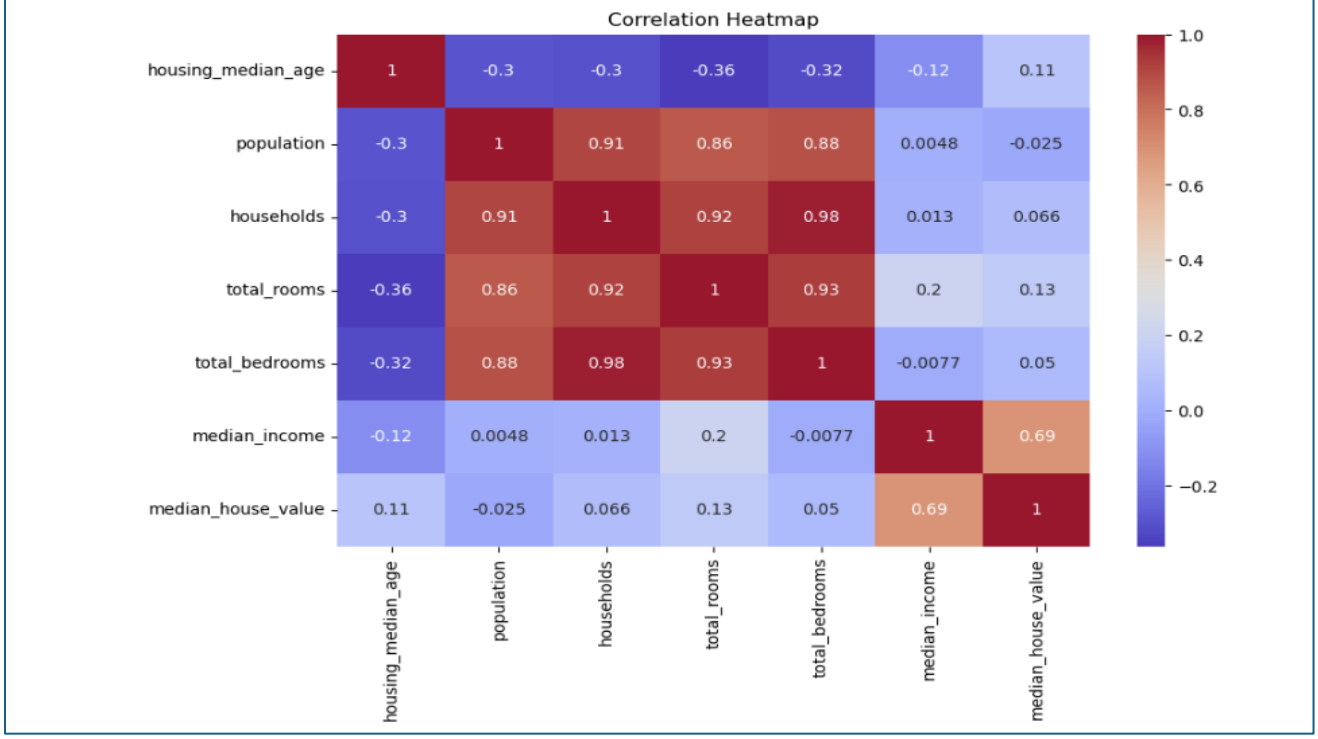

```
In [5]: op_counts = data['ocean_proximity'].value_counts()
explode=[0.01,0.01,0.01,0.01,0.01]
plt.pie(op_counts, labels=op_counts.index, autopct='%1.1f%%', explode=explode,startangle=90)
plt.title('House Ocean Proximity Distribution')
plt.show()
```



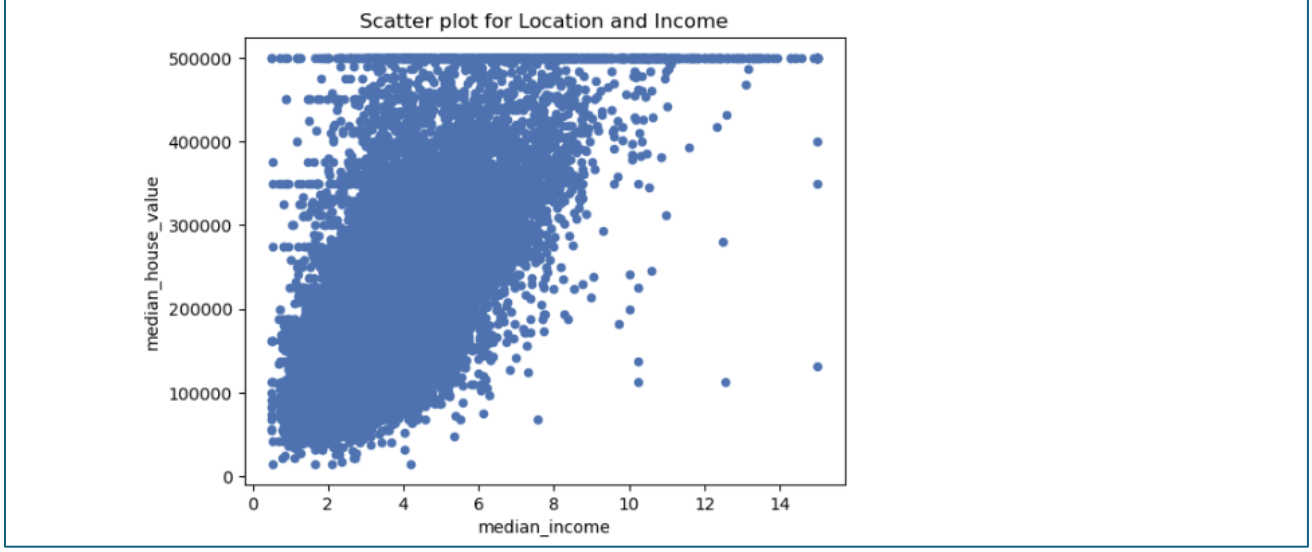
Most house in the dataset are from <1H from the ocean followed by houses inland. The Island number of houses is almost 0% of the dataset.

Chapter 3: Correlation between variables

```
In [20]:
# Heatmap for correlation
plt.figure(figsize=(10, 6))
sns.heatmap(data[['housing_median_age', 'population', 'households','total_rooms','total_bedrooms','median_income','median_hou
plt.title('Correlation Heatmap')
plt.show()
```



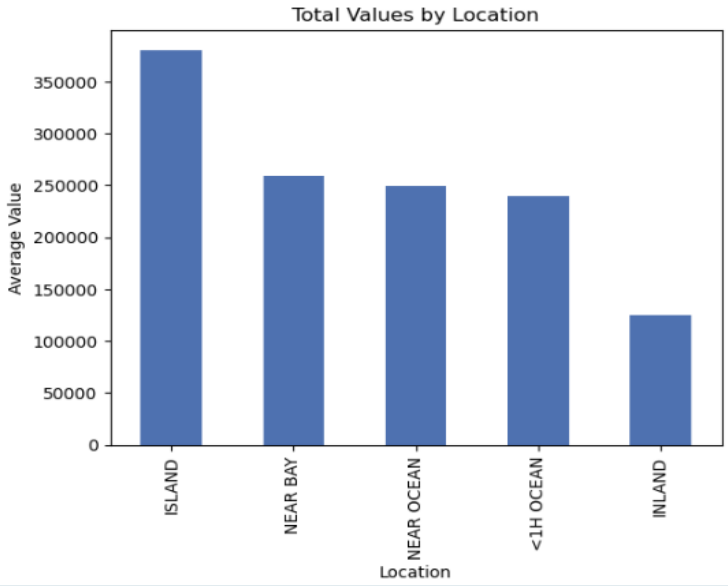

```
In [15]:
data.plot.scatter(x='median_income', y='median_house_value', title='Scatter plot for Location and Income');
```



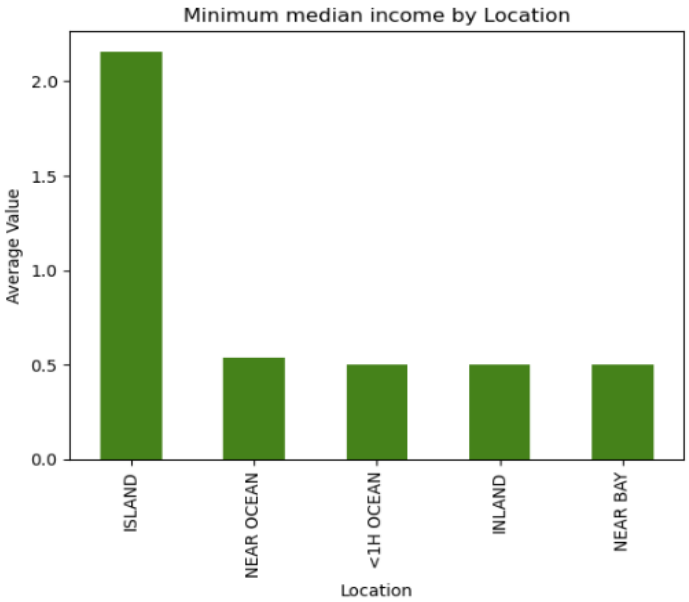
Strong positive correlation between population, household, total rooms and total bedrooms. Notable correlation between income and house value.

Chapter 4: House value by ocean proximity

In [38]:
```
data.groupby('ocean_proximity')['median_house_value'].mean().sort_values(ascending=False).head(10).plot(kind='bar')
plt.title('Total Values by Location')
plt.ylabel('Average Value')
plt.xlabel('Location')
plt.show();
```



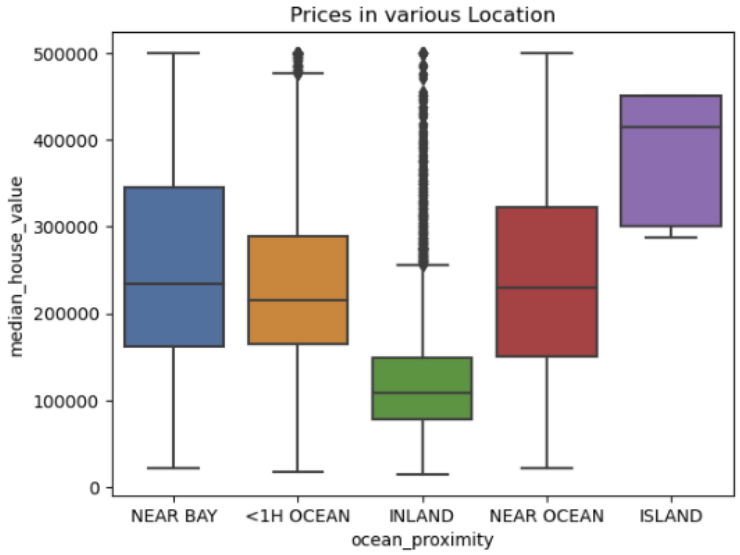
In [39]:
```
data.groupby('ocean_proximity')['median_income'].min().sort_values(ascending=False).head(10).plot(kind='bar', color ='g')
plt.title('Minimum median income by Location')
plt.ylabel('Average Value')
plt.xlabel('Location')
plt.show();
```



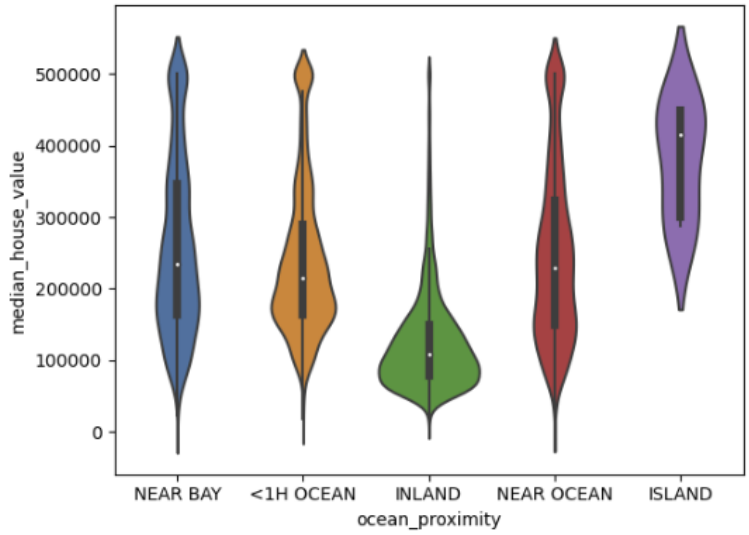
Houses in the island are more expensive than elsewhere with a minimum price being 4 times the minimum price elsewhere.

Chapter 5: Price Distribution by location

```
In [25]: sns.boxplot(x='ocean_proximity', y='median_house_value', data=data)
         plt.title('Prices in various Location');
```



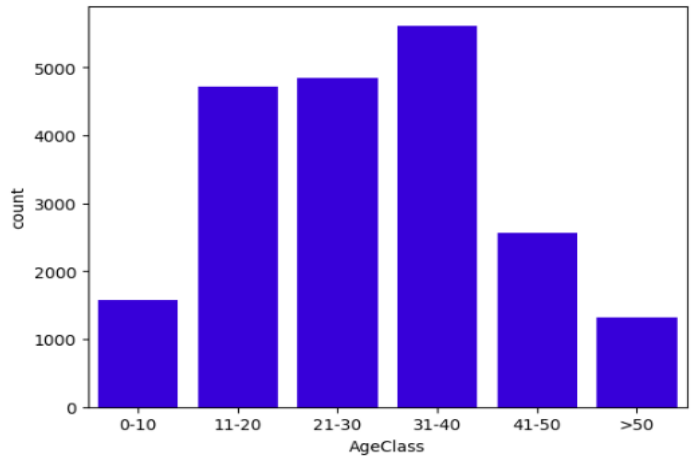

```
In [30]: sns.violinplot(x='ocean_proximity', y='median_house_value', data=data);
```



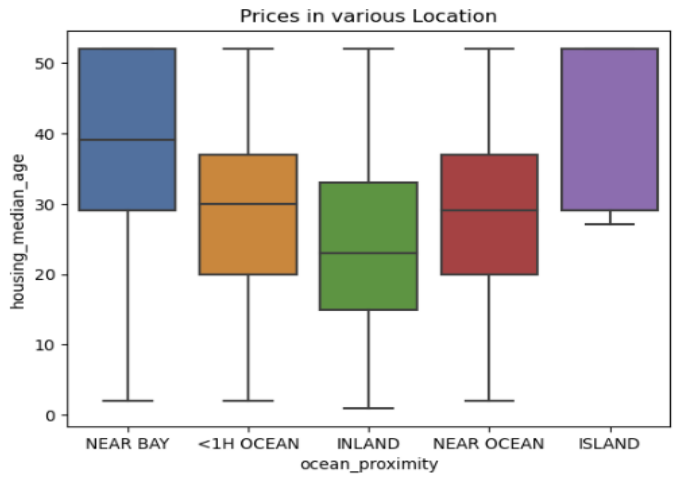
Houses are more expensive as you get closer to the ocean.

Chapter 6: House Age by ocean proximity

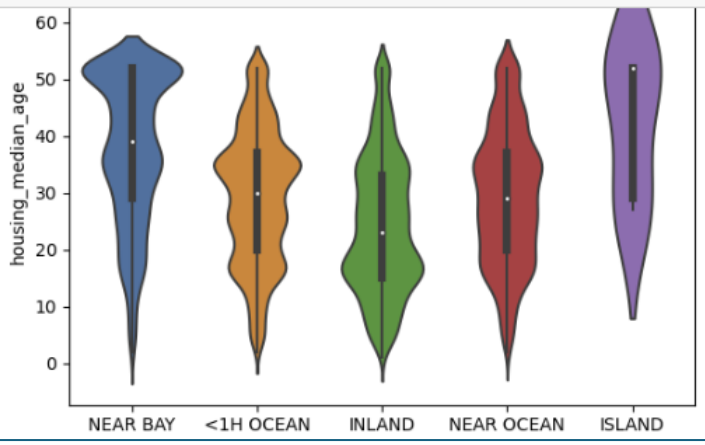
```
In [34]: sns.countplot(x='AgeClass', data=data, color = 'b')
Out[34]: <AxesSubplot:xlabel='AgeClass', ylabel='count'>
```



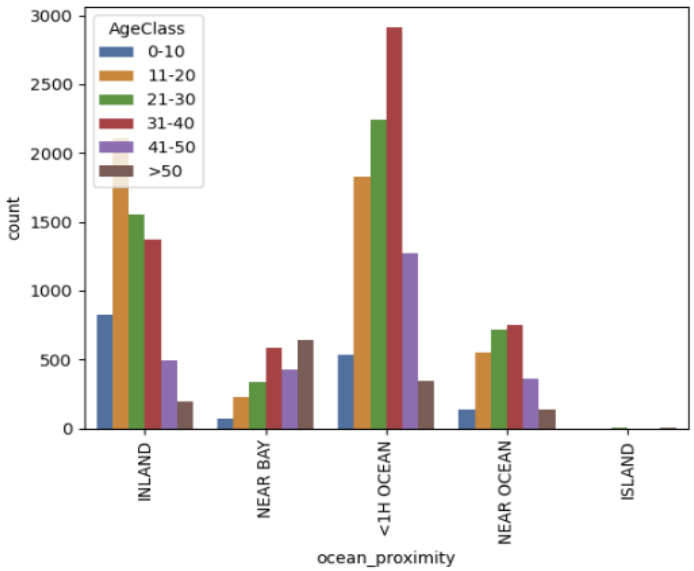

```
In [23]: sns.boxplot(x='ocean_proximity', y='housing_median_age', data=data)
         plt.title('Prices in various Location');
```



```
In [29]: sns.violinplot(x='ocean_proximity', y='housing_median_age', data=data);
```




```
In [33]: sns.countplot(x='ocean_proximity',hue='AgeClass', data=data)
         plt.xticks(rotation=90);
```



Houses are older as you get closer to the ocean.