California housing dataset

This colab introduces California Housing dataset that we will be using for regression demonstration.

We also list down the steps for typical dataset exploration, which can be applied broadly to any dataset.

Loading the dataset

This dataset can be fetched from `sklearn` with `fetch_california_housing` API.

```
from sklearn.datasets import fetch_california_housing
```

In order to analyze the dataset, let's load it as a dataframe.

```
california_housing = fetch_california_housing(as_frame=True)
```

```
type(california_housing)
```

```
sklearn.utils.Bunch
```

The bunch object is a dictionary like object with the following attributes:

- `data`, which is a pandas object (since `as_frame=True`). Each row corresponds to 8 feature values.
- `target` value contains average house value in units of 100,000. This is also a pandas object (since `as_frame=True`)
- `feature_names` is an array of ordered feature names used in the dataset.
- `DESCR` contains description of the dataset.
- `frame` contains dataframe with data and target.

Each of these attributes can be accessed as `<bunch_object>.key`. In our case, we can access these features as follows:

- `california_housing.data` gives us access to contents of `data` key.
- `california_housing.target` gives us access to contents of `target` key.
- `california_housing.feature_names` gives us access to contents of `feature_names` key.
- `california_housing.DESCR` gives us access to contents of `DESCR` key.
- `california_housing.frame` gives us access to contents of `frame` key.

Dataset exploration

STEP 1. Dataset description

Let's look at the description of the dataset

```
print(california_housing.DESCR)
```

```
.. _california_housing_dataset:

California Housing dataset
--------------------------

**Data Set Characteristics:**

    :Number of Instances: 20640

    :Number of Attributes: 8 numeric, predictive attributes and the target

    :Attribute Information:
        - MedInc        median income in block group
        - HouseAge      median house age in block group
        - AveRooms      average number of rooms per household
        - AveBedrms     average number of bedrooms per household
        - Population    block group population
        - AveOccup      average number of household members
        - Latitude      block group latitude
        - Longitude     block group longitude

    :Missing Attribute Values: None

This dataset was obtained from the StatLib repository.
https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html

The target variable is the median house value for California districts,
expressed in hundreds of thousands of dollars ($100,000).

This dataset was derived from the 1990 U.S. census, using one row per census
block group. A block group is the smallest geographical unit for which the U.S.
Census Bureau publishes sample data (a block group typically has a population
of 600 to 3,000 people).

An household is a group of people residing within a home. Since the average
number of rooms and bedrooms in this dataset are provided per household, these
columns may take surpinsingly large values for block groups with few households
and many empty houses, such as vacation resorts.

It can be downloaded/loaded using the
:func:`sklearn.datasets.fetch_california_housing` function.

.. topic:: References

    - Pace, R. Kelley and Ronald Barry, Sparse Spatial Autoregressions,
```

```
Statistics and Probability Letters, 33 (1997) 291-297
```

Note down key statistics from this description such as number of examples (or sample or instances) from the description.

- There are **20640 examples** in the dataset.
- There are **8 numerical attributes** per example.
- The target label is median house value.
- There are **no missing values** in this dataset.

STEP 2. Examine shape of feature matrix.

Number of examples and features can be obtained via `shape` of `california_housing.data`.

```
california_housing.data.shape
```

```
(20640, 8)
```

There are 20640 examples with 8 features as mentioned in the description.

```
type(california_housing.data)
```

```
pandas.core.frame.DataFrame
```

STEP 3. Examine shape of label

Let's looks at the shape of the label vector.

```
california_housing.target.shape
```

```
(20640,)
```

Target contains 20640 labels - one for each example.

```
type(california_housing.target)
```

```
pandas.core.series.Series
```

STEP 4. Feature names

Let's find out names of the attributes.

```
california_housing.feature_names
```

```
['MedInc',
 'HouseAge',
 'AveRooms',
 'AveBedrms',
 'Population',
 'AveOccup',
 'Latitude',
 'Longitude']
```

Note the attributes and their description, which is a key step in understanding the data.

- MedInc - median income in block
- HouseAge - median house age in block
- AveRooms - average number of rooms
- AveBedrms - average number of bedrooms
- Population - block population
- AveOccupancy - average house occupancy
- Latitude - house block latitude
- Longitude - house block longitude

STEP 5. Examine sample training examples

Let's look at a few training examples along with labels.

```
california_housing.frame.head()
```

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25

The dataset contains aggregated data about each district in California.

STEP 6. Examine features

Let's look at the features.

```
california_housing.data.head()
```

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25

We have information about

- Demography of each district(income, population, house occupancy),
- Location of the districts (latitude and longitude) and
- Characteristics of houses in the district (#rooms, #bedrooms, age of house)

Since the information is aggregated at the district levels, the features corresponds to averages or median.

STEP 7. Examine target

Let's look at the target to be predicted.

```
california_housing.target.head()
```

```
0    4.526
1    3.585
2    3.521
3    3.413
4    3.422
Name: MedHouseVal, dtype: float64
```

The target contains median of the house value for each district. You can see that the target is a real number and hence this is a regression problem.

STEP 8. Examine details of features and labels

Let's look at the details of features and target labels.

```
california_housing.frame.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   MedInc       20640 non-null  float64
 1   HouseAge     20640 non-null  float64
```

```
 2   AveRooms     20640 non-null  float64
 3   AveBedrms    20640 non-null  float64
 4   Population   20640 non-null  float64
 5   AveOccup     20640 non-null  float64
 6   Latitude     20640 non-null  float64
 7   Longitude    20640 non-null  float64
 8   MedHouseVal  20640 non-null  float64
dtypes: float64(9)
memory usage: 1.4 MB
```

We observe that

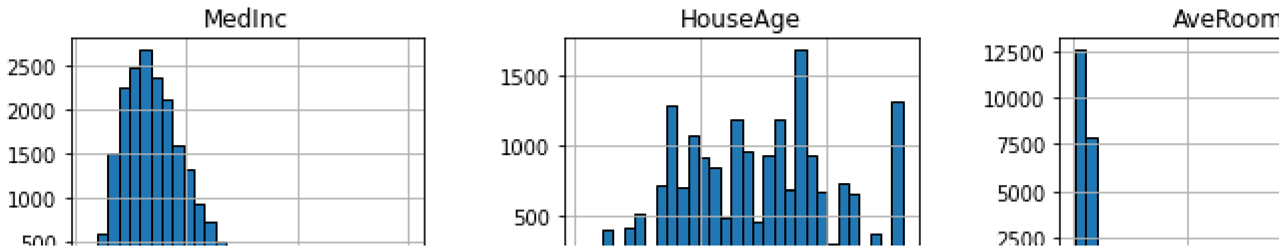
- The dataset contains 20640 examples with 8 features.
- All features are numerical features encoded as floating point numbers.
- There are no missing values in any features - the `Non-Null` is equal to the number of examples in the training set.

STEP 9. Feature and target histograms

Let's look at the distribution of these features and target by plotting their histograms.

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
california_housing.frame.hist(figsize=(12, 10), bins=30, edgecolor="black")
plt.subplots_adjust(hspace=0.7, wspace=0.4)
```



Let's observe these histogram and note down our findings:

- **MedInc** has a long tail distribution - salary of people is more or less normally distributed with a few folks getting a high salary.
- **HouseAge** has more or less a uniform distribution.
- The range for features, **AveRooms**, **AveBedrms**, **AveOccups**, **Population**, is large and it contains a small number of large values (as there are unnoticable bins on the right in the histogram plots of these features). That would mean that there could be certain outlier values present in these features.
- **Latitude** and **Longitude** carry geographical information. Their combination helps us decide price of the house.
- **MedHouseVal** also has a long tail distribution. It spikes towards the end. The reason is that the houses with price more than 5 are given value of 5.

STEP 10. Feature and target statistics

Let's look at statistics of these features and target.

```
california_housing.frame.describe()
```

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOc
count	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000
mean	3.870671	28.639486	5.429000	1.096675	1425.476744	3.070
std	1.899822	12.585558	2.474173	0.473911	1132.462122	10.386
min	0.499900	1.000000	0.846154	0.333333	3.000000	0.692
25%	2.563400	18.000000	4.440716	1.006079	787.000000	2.429
50%	3.534800	29.000000	5.229129	1.048780	1166.000000	2.818
75%	4.743250	37.000000	6.052381	1.099526	1725.000000	3.282
max	15.000100	52.000000	141.909091	34.066667	35682.000000	1243.333

We can observe that there is a large difference between `75%` and `max` values of `AveRooms`, `AveBedrms`, `Population` and `AveOccup` - which confirms our intuition about presence of outliers or extreme values in these features.

STEP 11. Pairplot

```
_ = sns.pairplot(data=california_housing.frame, hue="MedHouseVal", palette="viridis")
```

A few observations based on pairplot:

- `MedIncome` seems to be useful in distinguishing between low and high valued houses.
- A few features have extreme values.
- Latitude and longitude together seem to distinguish between low and high valued houses.

Summary

- Explored california housing dataset that would be used for demonstrating implementation of linear regression models.
- Examined various statistics of the dataset - #samples, #labels.
- Examined distribution of features through histogram and pairplots.