ĐẠI HỌC ĐÀ NẴNG

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG VIỆT - HÀN

VIETNAM - KOREA UNIVERSITY OF INFORMATION AND COMMUNICATION TECHNOLOGY

한-베정보통신기술대학교

Nhân bản – Phụng sự – Khai phóng

Chapter 2

Machine Learning Project

Machine Learning

CONTENTs

- Look at the Big Picture
- Get the Data
- Discover & Visualize the Data to Gain Insights.
- Prepare the Data for ML algorithms.
- Select & Train a Model
- Fine-Tune Model
- Present your solution.
- Launch, Monitor, & Maintain System.

• End-to-End Machine Learning Project

Main steps:

1. Look at the Big Picture
2. Get the Data
3. Discover & Visualize the Data to Gain Insights.
4. Prepare the Data for ML algorithms.
5. Select & Train a Model
6. Fine-Tune Model
7. Present your solution.
8. Launch, Monitor, & Maintain System.

- **Look at the Big Picture**
  - Build a model of housing prices in California using the California census data

⇨ What algorithms will be selected?

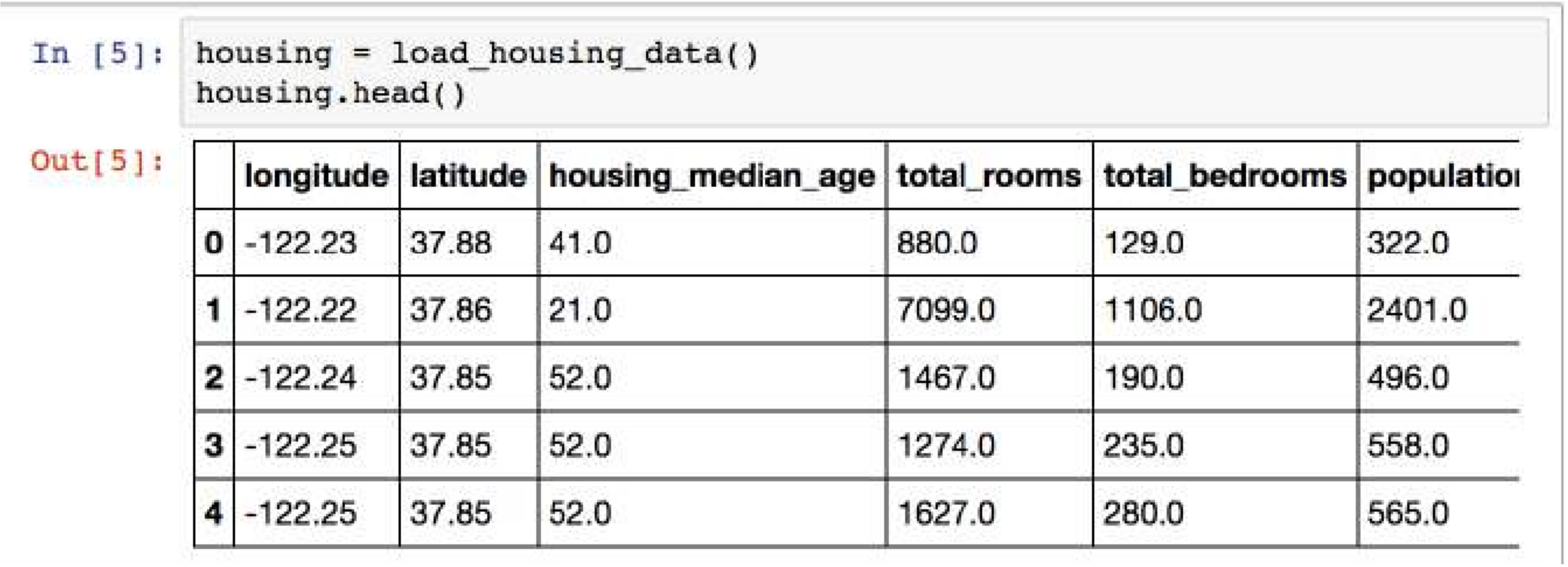
⇨ What performance measure will be used to evaluate the model?

⇨ How is this model used and benefit from it?

Select a Performance Measure:
the Root Mean Square Error (RMSE)

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2}$$

- **Get the Data**
  - Creating an Environment
  - Download the Data
  - Take a Quick Look at the Data Structure
  - Create a Test Set



```
In [5]: housing = load_housing_data()
        housing.head()
```

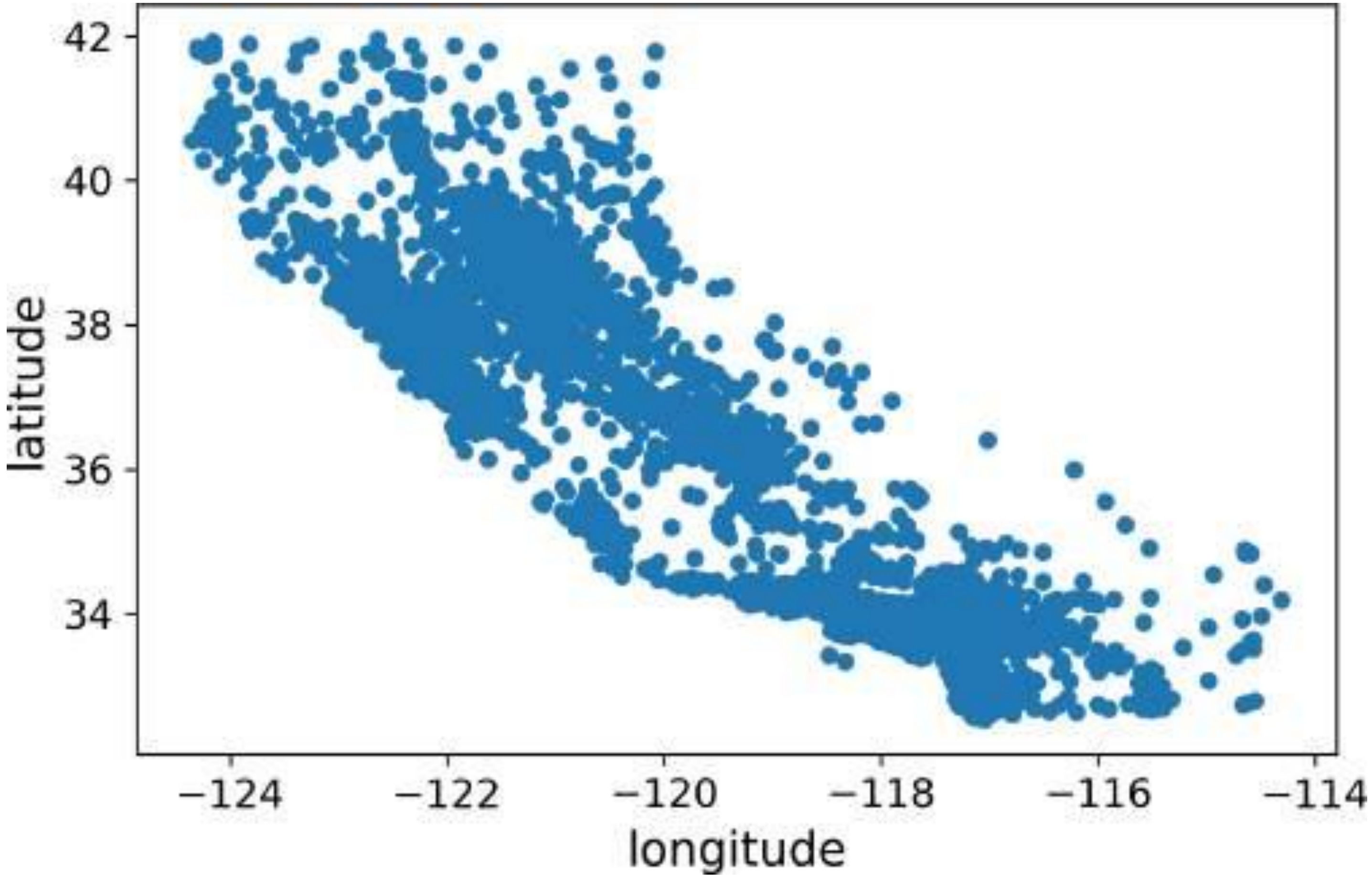
Out[5]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	populatio
0	-122.23	37.88	41.0	880.0	129.0	322.0
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0
2	-122.24	37.85	52.0	1467.0	190.0	496.0
3	-122.25	37.85	52.0	1274.0	235.0	558.0
4	-122.25	37.85	52.0	1627.0	280.0	565.0

Top five rows in the dataset

- **Discover & Visualize the Data to Gain Insights**

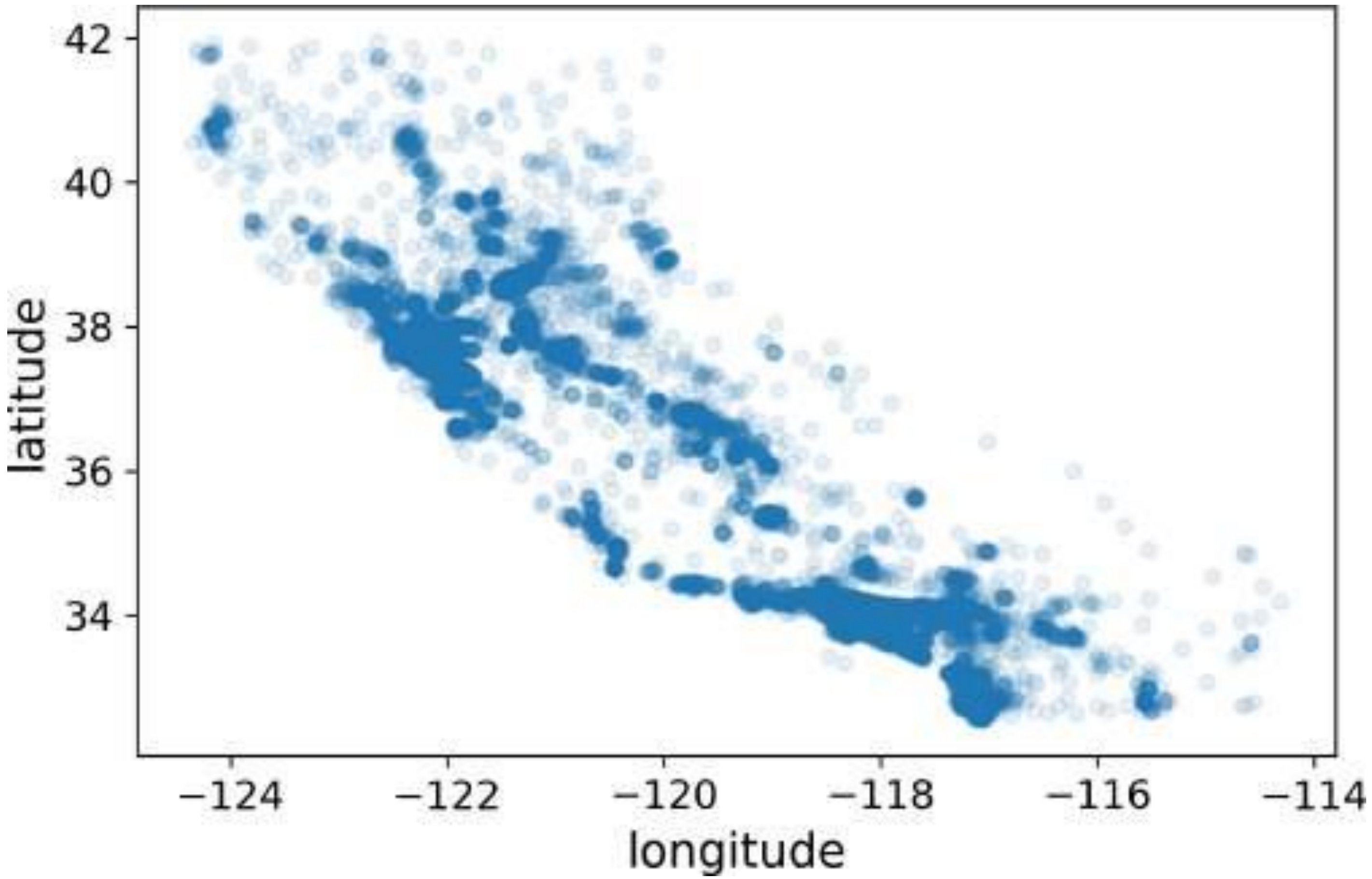
```
housing.plot(kind="scatter", x="longitude", y="latitude")
```



A geographical scatterplot of the data

- **Discover & Visualize the Data to Gain Insights**

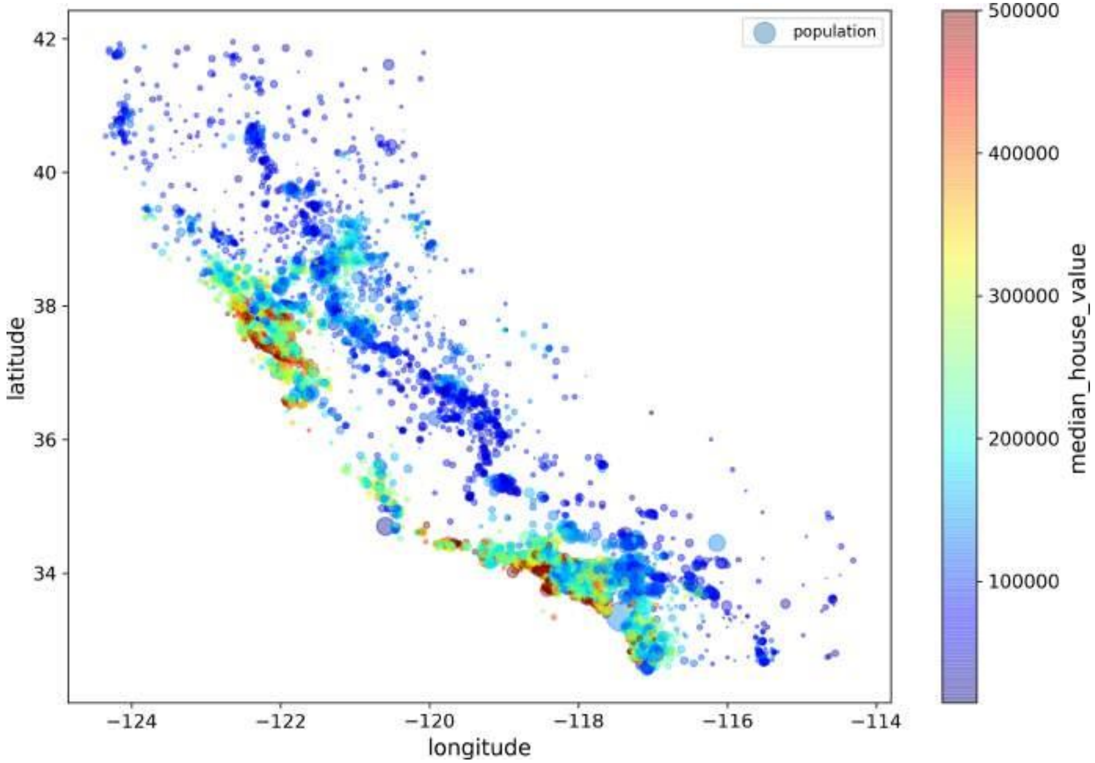
```
housing.plot(kind="scatter", x="longitude", y="latitude", alpha=0.1)
```



A better visualization highlighting high-density areas

- Discover & Visualize the Data to Gain Insights

```
housing.plot(kind="scatter", x="longitude", y="latitude", alpha=0.4,
    s=housing["population"]/100, label="population", figsize=(10,7),
    c="median_house_value", cmap=plt.get_cmap("jet"), colorbar=True,
)
plt.legend()
```



California housing prices

- **Prepare the Data for ML algorithms**
  - Data Cleaning
  - Handling Text and Categorical Attributes
  - Custom Transformers
  - Feature Scaling
  - Transformation Pipelines

- **Select and Train a Model**
  - Training and Evaluating on the Training Set
  - Better Evaluation Using Cross-Validation

- **Fine-Tune Model**
  - Grid Search
  - Randomized Search
  - Ensemble Methods
  - Analyze the Best Models and Their Errors
  - Evaluate Your System on the Test Set

- **Present your solution**
- **Launch, Monitor, & Maintain System**

SUMMARY

- Look at the Big Picture
- Get the Data
- Discover & Visualize the Data to Gain Insights
- Prepare the Data for ML algorithms
- Select & Train a Model
- Fine-Tune Model
- Present your solution
- Launch, Monitor, & Maintain System

Enjoy the Course…!