Urban Data Analysis

Final Project Preparation

Today's Agenda

- Selecting Datasets

- Defining Research Questions and Methodology

- Regression Case Study

- Final Project Expectations

Choosing Your Dataset

Good Datasets:

- Relevant to urban issues
- Sufficient size (1000+ records)
- Publicly available

Sources:

- NYC Open Data
- Data.gov
- FiveThirtyEight Datasets

Defining Research Questions

A Good Research Question:

- Specific and focused

- Quantitative and data-driven

- Feasible with available data

Example:

How does proximity to parks impact Brooklyn housing prices?

Planning Methodology

Decisions to Make:

- Independent Variables (X)

- Dependent Variable (y)

- Methods: Regression, clustering, visualization

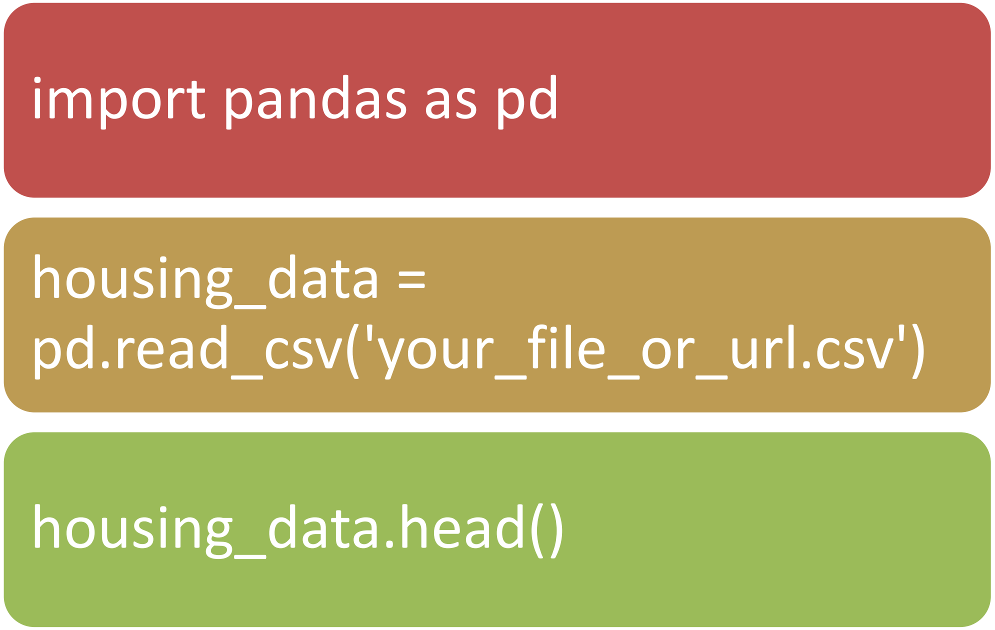
Case Study Setup

Scenario: Housing prices in NYC

Goal: Predict housing prices

Dataset: NYC Housing Sales

Step 1 - Load and Explore Data



```
import pandas as pd
```

```
housing_data = pd.read_csv('your_file_or_url.csv')
```

```
housing_data.head()
```

Step 2 - Clean and Prepare

Housing

```
housing_data = housing_data.dropna()
```

Housing

```
housing_data = housing_data[['SALE PRICE',
'GROSS SQUARE FEET', 'YEAR BUILT']]
```

Housing

```
housing_data =
housing_data.apply(pd.to_numeric,
errors='coerce').dropna()
```

Step 3 - Setup X and y

```
X = housing_data[['GROSS SQUARE FEET', 'YEAR BUILT']]

y = housing_data['SALE PRICE']

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Step 4 - Train a Model

```
from sklearn.linear_model
import LinearRegression
```

```
model = LinearRegression()
```

```
model.fit(X_train, y_train)
```

```
y_pred =
model.predict(X_test)
```

Step 5 - Evaluate Model

```
from sklearn.metrics import mean_squared_error, r2_score
```

```
print(mean_squared_error(y_test, y_pred))
```

```
print(r2_score(y_test, y_pred))
```

Step 6 - Visualize Results

```
import matplotlib.pyplot as plt
```

```
plt.scatter(y_test, y_pred)
```

```
plt.xlabel('Actual Sale Price')
```

```
plt.ylabel('Predicted Sale Price')
```

```
plt.title('Actual vs Predicted Housing Prices')
```

```
plt.show()
```

Final Project Expectations

- Select dataset by [next class date]

- Submit 3-sentence project proposal:

- Research Question

- Method(s)

- Target Variable

Discussion Questions

- How do you judge dataset quality?

- What makes a 'bad' regression model?

- How can you improve prediction accuracy?

Thank You!