Northeastern University
London

Lab: Programming with data

Lisias Loback

Welcome to Data Science Lab: Split-Apply-Combine Strategy

- What is Data Science
- Understand the Split-Apply-Combine Strategy
- Calculations on Data
- Hands-on Experience

Data Science

Data Science is a multidisciplinary field that combines various techniques from statistics, computer science, and domain expertise to extract valuable insights and knowledge from data.

Key Components

1. **Data Collection**
2. **Data Cleaning**
3. **Data Analysis**
4. **Data Visualization**
5. **Machine Learning**

Importance

1. **Informed Decision-Making**
2. **Improving Efficiency**
3. **Personalization**
4. **Predictive Analytics**
5. **Innovative Solutions**
6. **Competitive Advantage**

Understanding Split-Apply-Combine Method

1. Map-Reduce
2. Resampling
3. Pivoting and Melting
4. Window Functions
5. Vectorization
6. Filtering and Subsetting
7. Aggregations
8. Cross-Tabulation
9. Data Transformation
10. Data Integration

The Split-Apply-Combine method is a data analysis paradigm that involves splitting a dataset into groups, applying a function to each group independently, and then combining the results back into a single dataset. This method is particularly useful for aggregating data, computing summary statistics, and performing group-wise transformations.

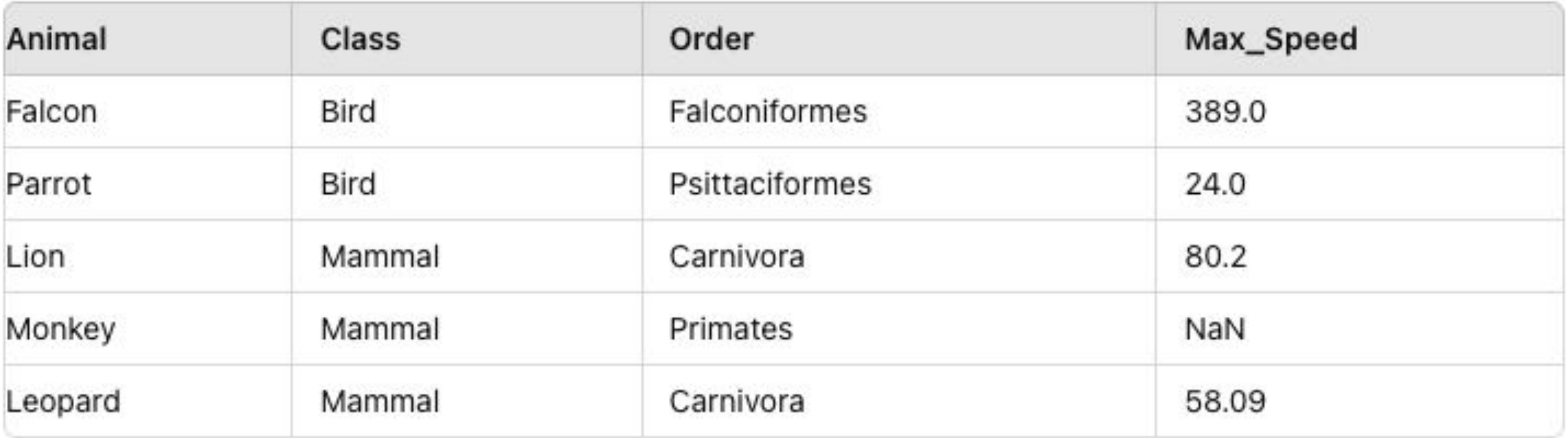
Sales Analysis

Customer Segmentation

Biological Research

Financial Reporting

Our Dataset: Animal Speeds



Animal	Class	Order	Max_Speed
Falcon	Bird	Falconiformes	389.0
Parrot	Bird	Psittaciformes	24.0
Lion	Mammal	Carnivora	80.2
Monkey	Mammal	Primates	NaN
Leopard	Mammal	Carnivora	58.09

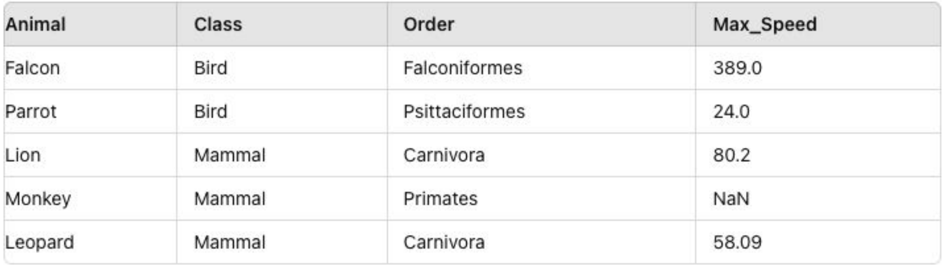
Our Dataset: Steps Applied

Split by Class:

- Birds: Falcon, Parrot
- Mammals: Lion, Monkey, Leopard

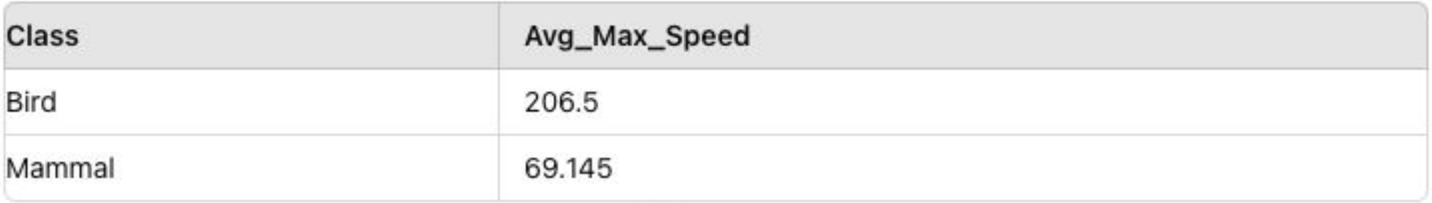
Apply Mean Function on Max_Speed:

- Birds: (389.0 + 24.0) / 2 = 206.5
- Mammals: (80.2 + 58.09) / 2 = 69.145



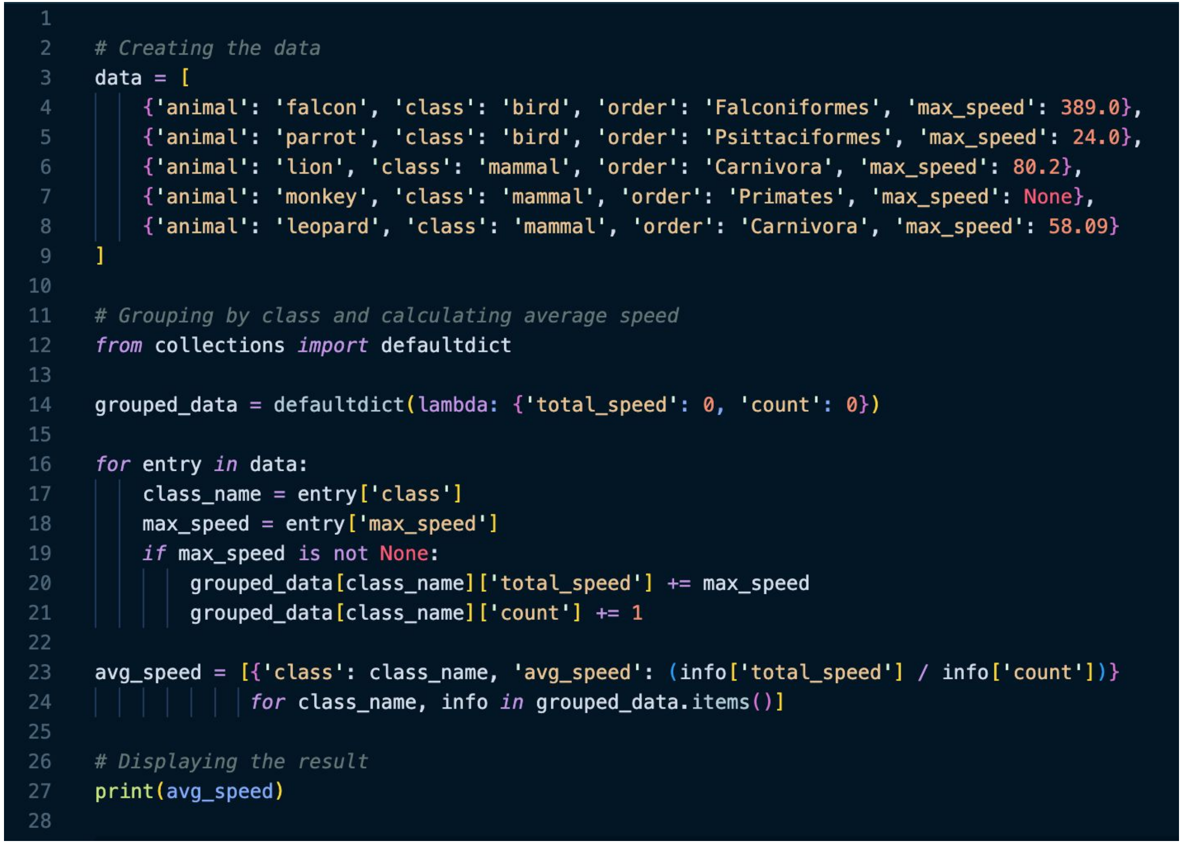
Animal	Class	Order	Max_Speed
Falcon	Bird	Falconiformes	389.0
Parrot	Bird	Psittaciformes	24.0
Lion	Mammal	Carnivora	80.2
Monkey	Mammal	Primates	NaN
Leopard	Mammal	Carnivora	58.09

Combine Results:

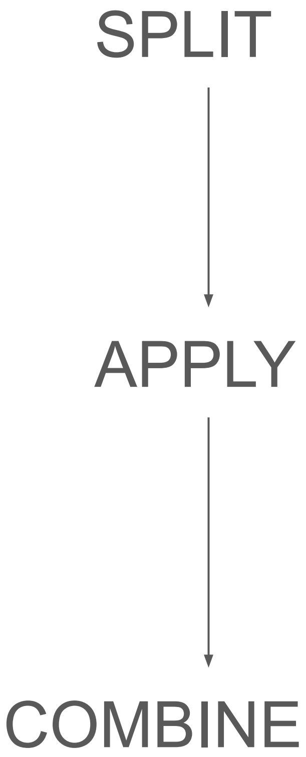


Class	Avg_Max_Speed
Bird	206.5
Mammal	69.145

Demonstration: Calculating Average Speeds



```python

# Creating the data
data = [
    {'animal': 'falcon', 'class': 'bird', 'order': 'Falconiformes', 'max_speed': 389.0},
    {'animal': 'parrot', 'class': 'bird', 'order': 'Psittaciformes', 'max_speed': 24.0},
    {'animal': 'lion', 'class': 'mammal', 'order': 'Carnivora', 'max_speed': 80.2},
    {'animal': 'monkey', 'class': 'mammal', 'order': 'Primates', 'max_speed': None},
    {'animal': 'leopard', 'class': 'mammal', 'order': 'Carnivora', 'max_speed': 58.09}
]

# Grouping by class and calculating average speed
from collections import defaultdict

grouped_data = defaultdict(lambda: {'total_speed': 0, 'count': 0})

for entry in data:
    class_name = entry['class']
    max_speed = entry['max_speed']
    if max_speed is not None:
        grouped_data[class_name]['total_speed'] += max_speed
        grouped_data[class_name]['count'] += 1

avg_speed = [{'class': class_name, 'avg_speed': (info['total_speed'] / info['count'])}
             for class_name, info in grouped_data.items()]

# Displaying the result
print(avg_speed)

```



SPLIT

APPLY

COMBINE

Calculating Average Speeds with pandas

Developer Environment: https://codesandbox.io/

Pandas: https://pandas.pydata.org/

https://pandas.pydata.org/docs/getting_started/index.html

https://www.w3schools.com/python/pandas/default.asp

https://www.w3schools.com/python/pandas/pandas_ref_dataframe.asp

Code:

https://github.com/lloback/labs/tree/main/split-apply-combine

Your Turn: Hands-on Practice

Load the dataset.

Perform the split-apply-combine method to calculate average max speeds.

Discuss the results with peers.

Q&A

Grus, J. (2019) *Data Science from Scratch: First Principles with Python*. O’Reilly Media.

Kotu, V. and Deshpande, B. (2018) *Data science: Concepts and Practice*. Morgan Kaufmann.