빅데이터와 인공지능의 이해

04.AI와 파이썬 라이브러리 활용

조창제

강의자료

2025.03.

목차

CONTENTS

목차

CONTENTS

1.라이브러리와 패키지

1) 라이브러리

① 특정 작업을 수행하기 위해 제공되는 재사용 가능한 코드 집합(함수, 클래스, 모듈 등)

2) 패키지

① 관련된 여러 모듈과 하위 라이브러리들의 집합

3) 파이썬

- 파이썬 공식 저장소 PyPI(Python Package Index) 저장소 기준 **530,000개 이상**의 라이브러리가 등록되어 있음
- 라이브러리를 완벽히 공부하기보다는 **필요한 라이브러리 위주로 모듈을 응용하는 방법을 악하는 것이 효과적**

Library
(공구 하나)

Package
(공구 집합)

01.numpy

1.정의

1) 파이썬의 고성능 수치 계산을 위한 라이브러리

① 다양한 자료타입과 자료 생성 함수를 지원

자료타입	설명
np.int64	정수
np.float64	실수
np.complex	복소수
np.bool	논리(True, False)
np.object	객체구조
np.string_	고정길이 문자열
np.unicode_	유니코드 문자 저장 가능

자료생성	설명
np.array	배열을 생성
np.zeros	0값을 가진 numpy 객체 생성
np.ones	1값을 가진 numpy 객체 생성
np.full	특정 값을 가진 numpy 객체 생성
np.eye	단위행렬 생성
np.linspace	start, end까지 n개의 선형적 값 생성
np.random.random	0~1사이 실수 생성
np.empty	메모리 상의 이전값들로 채워진 배열 생성

1.정의

1) 파이썬의 고성능 수치 계산을 위한 라이브러리

② 다양한 수학적 기능을 제공

- 집계함수 지원

함수	설명
np.add(a,b)	더하기
np.subtract(a, b)	빼기
np.multiply(a, b)	곱하기
np.divide(a, b)	나누기
np.exp(a)	지수연산
np.sqrt(a)	제곱근
np.log(a)	로그
np.sin(a), np.cos(a), np.tan(a)	삼각함수
a.dot(b), @	행렬곱

함수	설명
a.sum, a.cumsum	합계, 누적합
a.min, a.max	최소값, 최댓값
a.mean, a.median, a.std	평균, 중앙값, 표준편차
a.corrcoef	상관계수
np.diag	주대각성분
np.linalg.det	행렬식

01.numpy

1.정의

1) 파이썬의 고성능 수치 계산을 위한 라이브러리

③ 다양한 배열의 처리에 특화

- 데이터 형태변환 및 자료 추출/결합 지원

함수	설명
np.sort	정렬
a[0:2], a[1,...], a[:1], a[...,-4:]	슬라이싱
a[2], a[1,2]	부분추출
a[a>2], a[[1,3,0],[1,2]]	인덱싱
a.flatten	객체구조
a.T, np.transpose(a)	전치
a.reshape	구조변형
np.concatenate	결합
np.hstack, np.vstack	수평/수직 결합
np.hsplit, np.vsplit	수평/수직분해

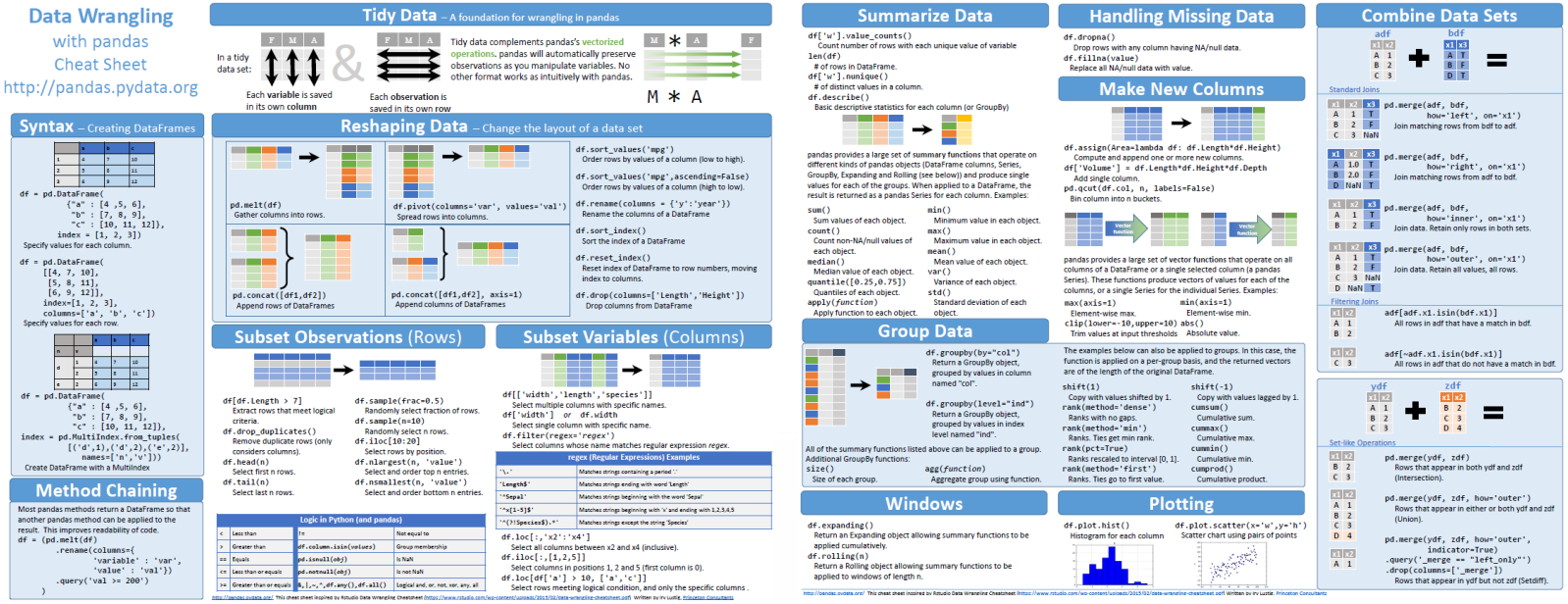
02.pandas

1.정의

1) 표 형식의 데이터를 다루는데 유용한 라이브러리

① 데이터 변환, 처리, 분석에 필요한 다양한 기능을 제공

- 다양한 데이터(CSV, Excel, SQL, JSON 등) 저장 및 불러오기 가능
- 데이터 탐색 및 필터링 가능, 데이터 변환 및 처리, 정렬 가능
- 결측값 처리 가능, 그룹화 및 집계 가능



2.활용

1) pandas 객체 추출

① iloc(Integer location): 정수 위치 기반 인덱싱
② loc(Label based location): 레이블 기반 인덱싱

객체 추출 관련 함수	설명
df.sample	데이터 프레임 내에서 랜덤 샘플 추출
df.dropna	결측 데이터를 제거
df.fillna	결측 데이터를 보간
df.drop_duplicates	중복 데이터를 제거
df.head	위에서부터 데이터 프레임 일부 추출
df.tail	아래에서부터 데이터 프레임 일부 추출

객체 변환 관련 함수	설명
df.melt	열을 행으로 변환
df.pivot	특정 열의 값을 새로운 열로 변환
pd.concat	데이터프레임을 행/열 기준으로 병합
pd.merge	공통 열 기준으로 데이터프레임 병함

집계 관련 함수	설명
df[[columns]].value_counts	컬럼들 기준으로 값의 개수를 카운팅
df.groupby	특정 열을 기준으로 그룹화

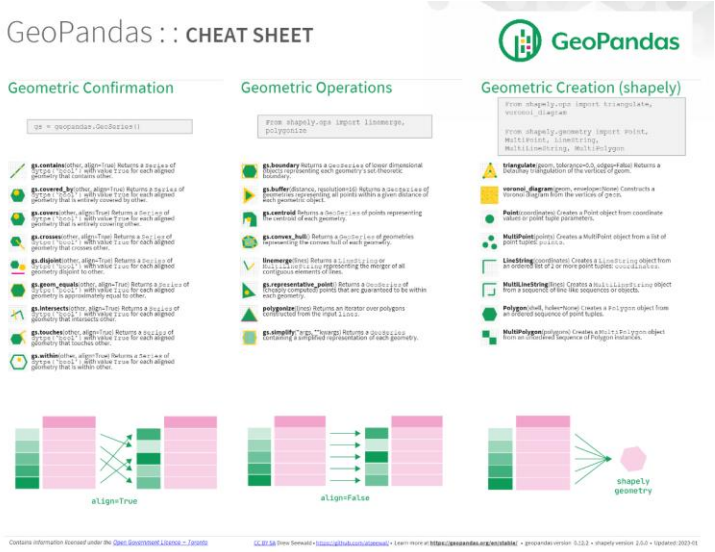
03.geo-pandas

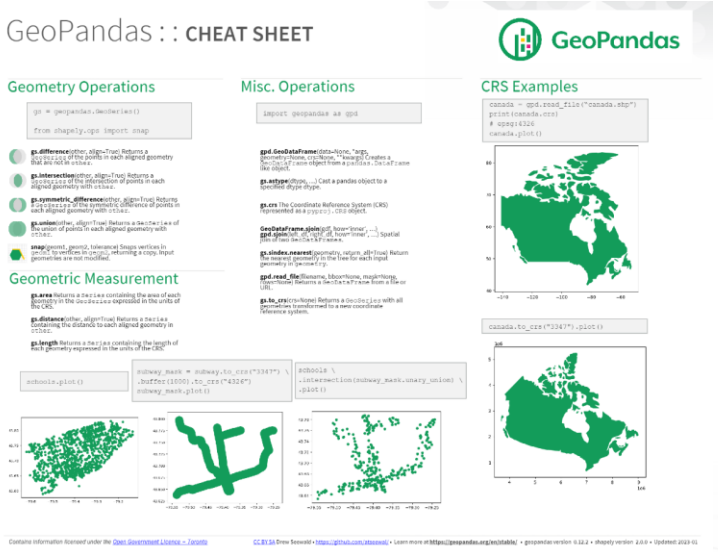
1.정의

1) 공간 데이터를 처리하고 분석하기 위한 라이브러리

① 공간 데이터를 생성하거나 저장 가능

② 공간 데이터에 대한 결합, 분할, 추출 등이 가능

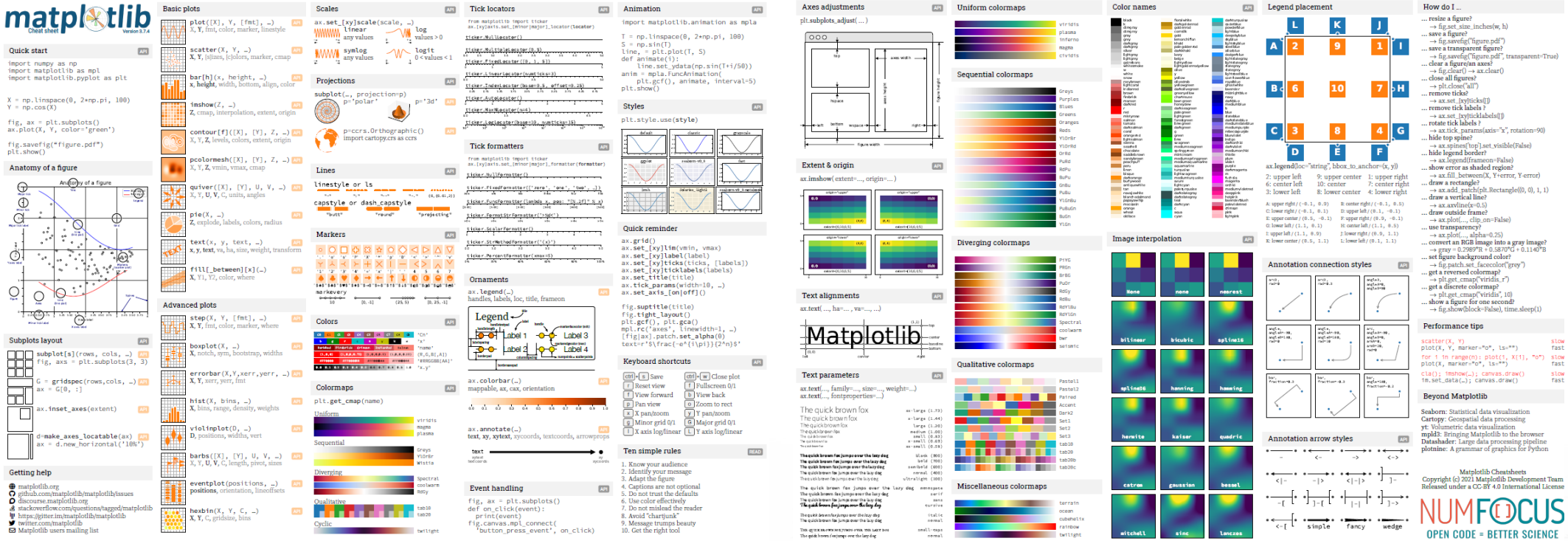




01.matplotlib

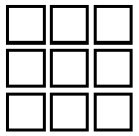
1.정의

1) python의 데이터 시각화를 위한 라이브러리
2) 시각화한 자료를 다양한 확장자로 저장이 가능

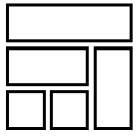


2.활용

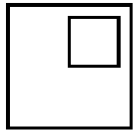
1) 한 그림에 여러 그래프를 그릴 수 있음
2) 여러 그래프를 배열하여 그릴 수도 있음
3) 그림의 여백을 상세하게 수정할 수 있음



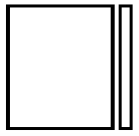
```
subplot[s](rows, cols, …)  API
fig, axs = plt.subplots(3, 3)
```



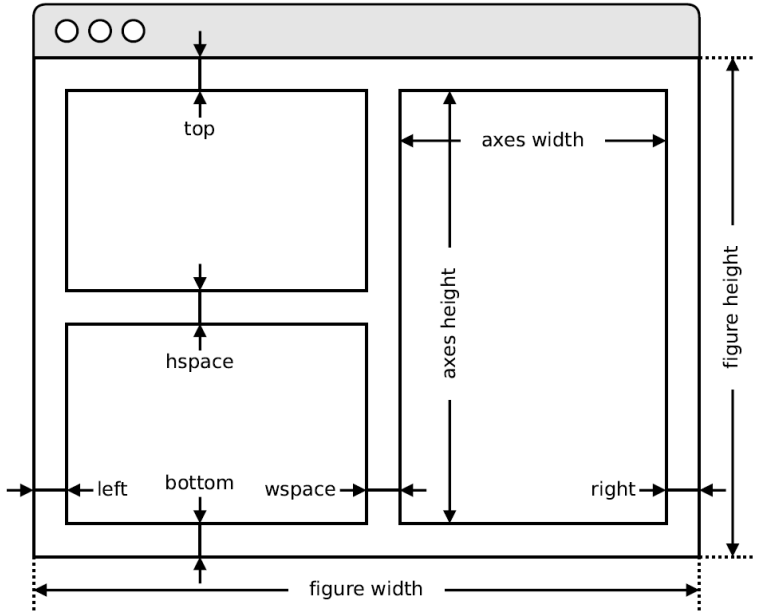
```
G = gridspec(rows,cols, …)  API
ax = G[0, :]
```

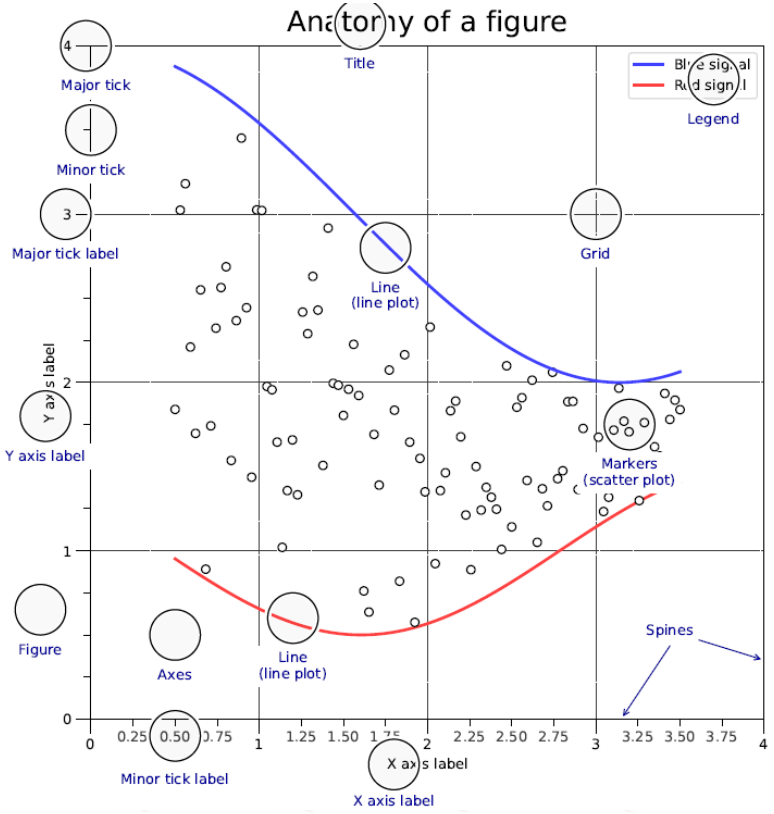


```
ax.inset_axes(extent)  API
```



```
d=make_axes_locatable(ax)  API
ax = d.new_horizontal('10%')
```





01.scikit-learn

1.정의

1) python의 머신러닝 패키지 중 하나

① 데이터 전처리

- StandardScaleler, MinMaxScaler, RobustScaler
- OneHotEncoder, LabelEncoder

② 데이터 분할

- train_test_split, Kfold, LeaveOneOut
- StratifiedShuffleSplit, StratifiedKFold: 특정 클래스 비율 유지
- GrouShuffleSplit, GroupKFold: 특정 그룹(환자ID, 사용자ID 등)을 기준으로 랜덤하게 분할

01.scikit-learn

1.정의

1) python의 머신러닝 패키지 중 하나

① 다양한 모델 지원(지도, 비지도 학습)

- 지도학습 – 분류/회귀
  - 트리기반 – DecisionTree, RandomForest, GradientBoosting, XGB, LGB, AdaBoost
  - 그 외 – Logistic Regression, Linear Regression, Lasso, Ridge, SVM, LDA, QDA
- 비지도학습
  - 차원축소 – PCA, SVD, tSNE, MDS
  - 클러스터링 – Kmeans, DBSCAN 등
- 연관규칙 – Apriori

② 모델 학습 및 하이퍼 파라미터 튜닝

③ 모델 평가 기능 지원

Thank You

Email: qkdrk777777@naver.com