3. Python 기초 - Pandas

전주정보문화산업진흥원
Jeonju IT&CT Industry Promotion Agency

Data Insight

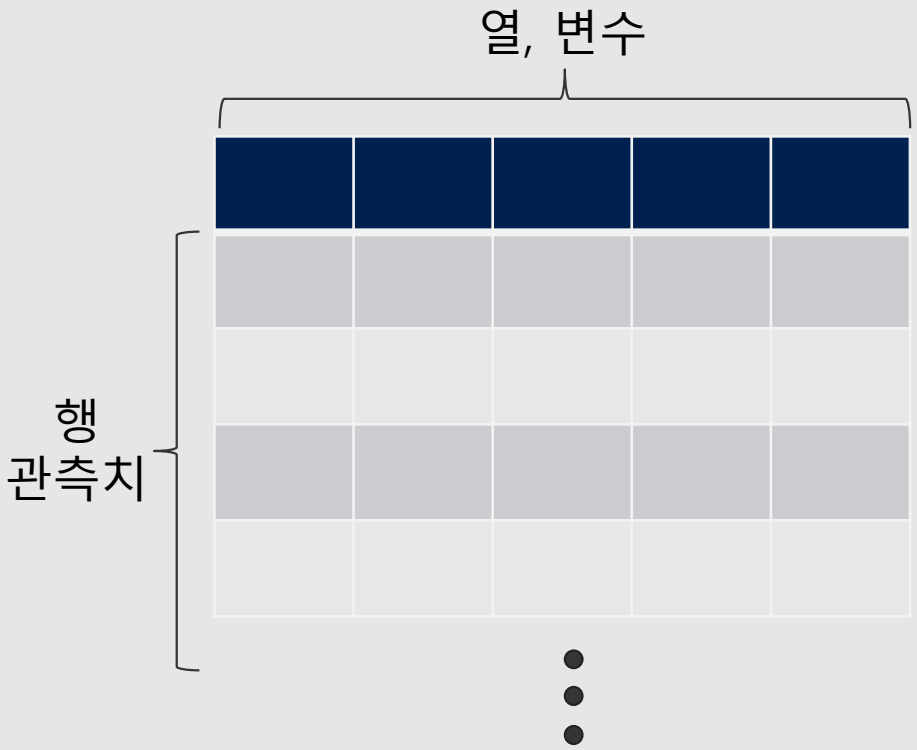
순서

- ✓ Dataframe 생성
- ✓ CSV 파일에서 데이터 불러오기
- ✓ 데이터 미리보기
- ✓ 원하는 데이터 조회하기
- ✓ Dataframe 수정하기
- ✓ Group by
- ✓ Merge

Dataframe 생성

✓Dataframe이란?

- 데이터 분석에서 가장 중요한 데이터구조
- RDBMS에서의 테이블 형태
- 변수들의 집합
  → 각 열을 변수라고 부른다!
- 행렬과 다른 점은?



Dataframe 생성

✓ dp.Dataframe(dirctionary형태)

	col1	col2	col3
0	Item0	Gold	1
1	Item0	Bronze	2
2	Item1	Gold	3
3	Item1	Silver	4

Python

```
# Dataframes 생성
d = {
  'col1': ['Item0', 'Item0', 'Item1', 'Item1']
 ,'col2': ['Gold', 'Bronze', 'Gold', 'Silver']
 ,'col3': [1, 2, 3, 4]
}
df = pd.DataFrame(d)
print(df)
```

CSV파일에서 데이터 불러오기

- ✓ 데이터를 가져오고자 할 때
  - Database에서 직접 가져오거나
  - CSV 파일에서 데이터를 불러온다.
- ✓ pd.read_csv( )
- ✓ pd.to_csv( )

Python

```
# Loading CSV files
df = pd.read_csv('Graduate_apply.csv', sep=',')

print(df.head())

df = pd.read_csv('Graduate_apply.csv'
                 , sep=','
                 , skipinitialspace=True)
print(df.head())

# to_csv
df.to_csv('./file.csv', header=True, index=False
          , encoding='utf-8')
```

Pandas DataType

Pandas dtype	Python type	NumPy type	Usage
object	str	string_, unicode_	Text
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float_, float16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values

데이터 살펴 보기①

- ✓상위, 하위 데이터 조회
  - df.head(#) : #값이 없으면 default는 5
  - df.tail(#) : #값이 없으면 default는 5
- ✓데이터프레임 모양 확인
  - df.shape
- ✓칼럼명들 조회
  - Df.columns

Python

```
# 첫 5개 행의 데이터를 보여줍니다.
df.head()

# 마지막 3개 행의 데이터를 보여줍니다.
df.tail(3)

# 데이터 프레임 모양 확인
df.shape

# 칼럼명 출력
Print(df.columns)
```

데이터 살펴 보기②

- ✓ 기초통계량
  - df.describe()
- ✓ Sorting
  - df.sort_values(), df.sort_index()

Python

```
# 간단한 통계 정보, 기초통계량,
df.describe()

# index로 정렬
df.sort_index(axis=0, ascending=False).head()

# 특정 컬럼의 값으로 정렬
df.sort_values(by=['admit', 'gpa’]
                          , ascending=False).head()
```

실습 #2 : csv 파일 불러와서 살펴 보기

원하는 데이터 조회하기 ①

✓ 칼럼명으로 조회

- df['칼럼명'] : Series로 결과 출력
- df[['칼럼명']] : Dataframe으로 결과 출력

Python

```
# 칼럼명으로 조회 1
df['gre'].head()
df['gre'].unique()

# 칼럼명으로 조회 2
df[['gre']].head()

# 두 개의 칼럼 동시 조회
df[['gpa', 'gre']].head()
```

원하는 데이터 조회하기 ②

✓Index로 조회

- df.iloc[row, column]

Python

```
# 행번호 1~3 조회
df.iloc[1:3]

# 0~4 rows & 0~2 columns
df.iloc[0:4, 0:2]
```

원하는 데이터 조회하기 ③

✓조건으로 조회

- df.loc[row조건, col조건]
- ==, !=, >=, <=, >, <
- .isin([val1, val2, ...])
- &(and), |(or)
- .str.contains(문자열)

Python

```
# Query by a single column value
df[df['gpa'] > 3.0].head()

# in 연산자
df[df['rank'].isin([1, 2])].head()

# &(and), |(or) 연산
df[(df['gpa'] > 3.0) & (df['rank'] == 3)].head()

df[(df['gpa'] > 3.0) | (df['rank'] == 3)].head()

# 문자열 포함하는 행 조회
df1[df1.col2.str.contains('ilver')]
```

실습 #3 : 원하는 데이터 조회하기

Dataframe 수정하기

✓ Dummy Variable

- 범주형 데이터를 숫자로 변환하기

계절	
봄	1
여름	2
가을	3
겨울	4

봄	여름	가을	겨울
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

✓ 열 단위로 합치기

- 범주형 데이터를 숫자로 변환하기

Python

```
# 특정 칼럼의 Dummy Variable을 얻기
df_rank = pd.get_dummies(df['rank'])
print(df_rank.head())
print("------")

# Dummy 데이터를 원래 데이터와 합치기
df_new = pd.concat([df, df_rank], axis=1)
print(df_new.head())
print("------")
```

Dataframe 열 제거하기

- ✓ df.drop()
  - axis=1 : 칼럼을 삭제
  - inplace=True : df에 직접 삭제

Python

```
# 특정 칼럼 제거하기
df_new.drop('rank', axis=1, inplace=True)
print(df_new.head())

# 여러 칼럼 동시 제거하기
df_new.drop(['gre', 'gpa'], axis=1, inplace=True)
print(df_new.head())
```

실습 #4 : 데이터프레임 수정하기

Group by

✓ Group by
- 특정 열 기준으로 연속형 값 집계

Python

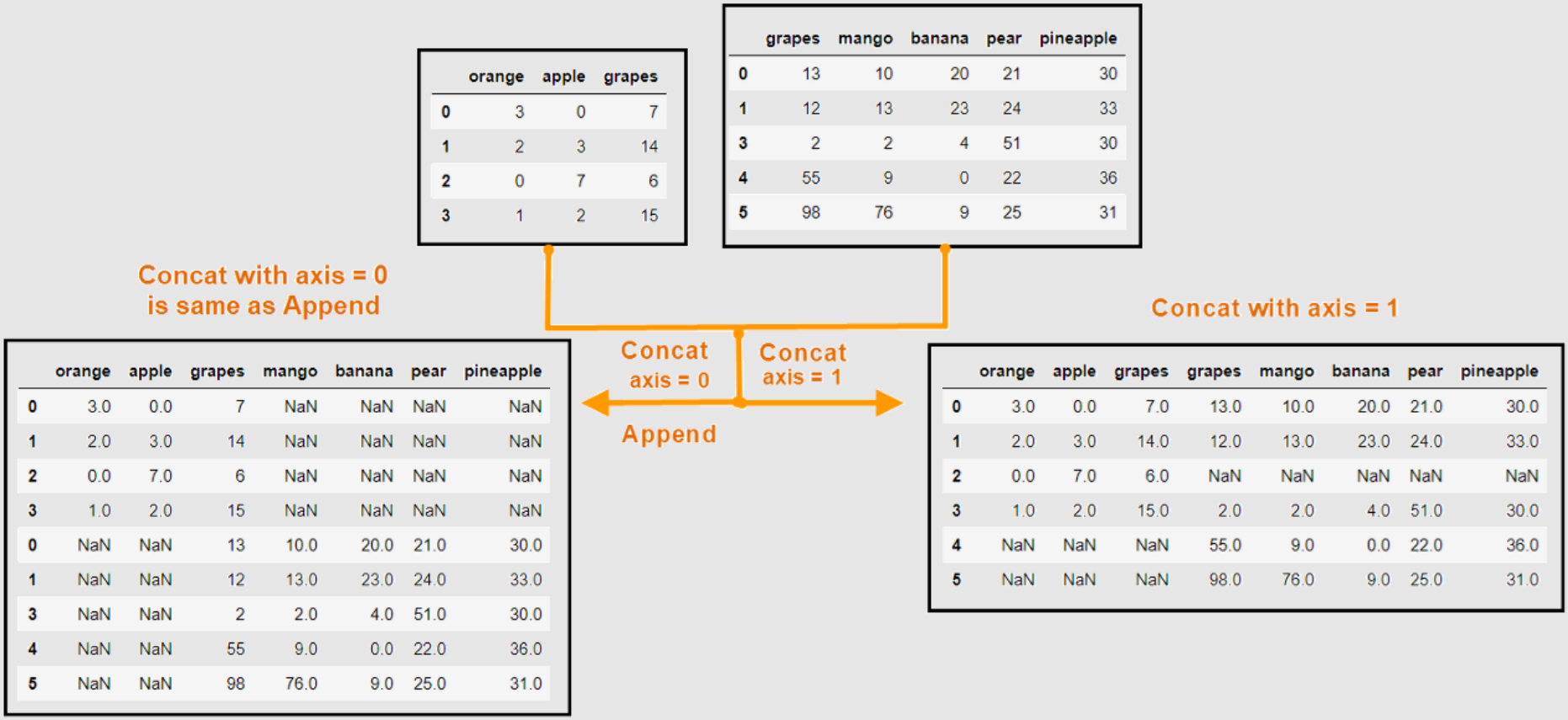
```
# rank별 평균 gre 값을 구하시오.
df.groupby(by=['rank'], as_index = False)['gre'].mean()

#as_index = True1이면 결과가 series로 나옴.
df.groupby(by = ['rank', 'admit']
                        , as_index=False)['gre', 'gpa'].mean()
```

실습 #5 : Group by

데이터 프레임 결합 : concat

✓두 데이터 프레임을 열로(옆으로), 행으로(위,아래로) 붙이기



	orange	apple	grapes
0	3	0	7
1	2	3	14
2	0	7	6
3	1	2	15

	grapes	mango	banana	pear	pineapple
0	13	10	20	21	30
1	12	13	23	24	33
3	2	2	4	51	30
4	55	9	0	22	36
5	98	76	9	25	31

	orange	apple	grapes	mango	banana	pear	pineapple
0	3.0	0.0	7	NaN	NaN	NaN	NaN
1	2.0	3.0	14	NaN	NaN	NaN	NaN
2	0.0	7.0	6	NaN	NaN	NaN	NaN
3	1.0	2.0	15	NaN	NaN	NaN	NaN
0	NaN	NaN	13	10.0	20.0	21.0	30.0
1	NaN	NaN	12	13.0	23.0	24.0	33.0
3	NaN	NaN	2	2.0	4.0	51.0	30.0
4	NaN	NaN	55	9.0	0.0	22.0	36.0
5	NaN	NaN	98	76.0	9.0	25.0	31.0

	orange	apple	grapes	grapes	mango	banana	pear	pineapple
0	3.0	0.0	7.0	13.0	10.0	20.0	21.0	30.0
1	2.0	3.0	14.0	12.0	13.0	23.0	24.0	33.0
2	0.0	7.0	6.0	NaN	NaN	NaN	NaN	NaN
3	1.0	2.0	15.0	2.0	2.0	4.0	51.0	30.0
4	NaN	NaN	NaN	55.0	9.0	0.0	22.0	36.0
5	NaN	NaN	NaN	98.0	76.0	9.0	25.0	31.0

데이터 프레임 결합 : Merge

✓특정 열 기준으로 두 데이터 프레임 붙이기

	key	Name	Age
0	K0	Jai	27
1	K1	Princi	24
2	K2	Gaurav	22
3	K3	Anuj	32

	key	Address	Qualification
0	K0	Nagpur	Btech
1	K1	Kanpur	B.A
2	K2	Allahabad	Bcom
3	K3	Kannuaj	B.hons

	key	Name	Age	Address	Qualification
0	K0	Jai	27	Nagpur	Btech
1	K1	Princi	24	Kanpur	B.A
2	K2	Gaurav	22	Allahabad	Bcom
3	K3	Anuj	32	Kannuaj	B.hons

실습 #6 : Merge

✓ 세 테이블(데이터프레임)으로 연습해 봅시다.

Products

	ProductID	ProductName	Category	SubCategory
0	p1052661	새우깡	간식	과자
1	p1054261	고구마스틱	간식	과자
2	p1097821	짱구	간식	과자
3	p1097831	감자칩	간식	과자

Customers

	CustomerID	RegisterDate	Address	Gender	BirthYear	Addr1	Addr2
0	c328222	2014-09-25	강원 원주시 늘품로	F	1960	강원도	원주시
1	c281448	2013-06-18	강원 원주시 치악로	F	1974	강원도	원주시
2	c038336	2003-10-10	강원 춘천시 서부대성로	F	1968	강원도	춘천시
3	c084237	2007-03-09	강원도 강릉시 연곡면 황어대길	F	1982	강원도	강릉시
4	c162600	2010-06-14	강원도 속초시 농공단지길	F	1978	강원도	속초시

Sales

	OrderID	Seq	OrderDate	ProductID	Qty	Amt	CustomerID
0	107	2	2016-01-02	p1036481	2	2100	c150417
1	69	1	2016-01-02	p1152861	1	1091	c212716
2	69	7	2016-01-02	p1013161	1	2600	c212716
3	69	8	2016-01-02	p1005771	1	1650	c212716