분석 틀

2024 빅데이터 분석 중간고사 대비 - 나만의 분석 틀

1. 파일 백업(저장하기)

```
df = pd.read_csv(" ", index_col='열') #파일 불러오기
df.to_csv("save.csv") # 파일 저장하기

check_file = pd.read_csv('save.csv")
check_file # 파일이 제대로 백업 되었는 지 확인
```

2. 데이터 확인하기

```
# 데이터 구성 확인해보기
df.column  # 데이터의 컬럼을 확인해본다.
df.values  # 데이터의 구성을 알아본다.
df.index

df.max() # 각 열의 연속형 데이터 최댓값을 확인하면서 이상치도 같이 확인해
df.min() # 각 열의 연속형 데이터 최솟값을 확인하면서 이상치도 같이 확인해

df.isnull().sum() # 데이터에 널이 있는지 확인하기
df.duplicated()   # 데이터에 중복된 값이 있는지 확인한다.

df.colums, df.values, df.max(), df.min(), df.isnull().sum(),

# 결측치가 있는 것을 확인했다. 결측치가 있는 열의 특성을 생각했을 때,
# 해당 열의 평균으로 대체하는것이 가장 좋다고 판단이 된다.
# 고로 결측치를 해당 열의 평균값으로 대체한다.
# 하지만 중복된 값과 이상치가 있는 것을 확인했으므로
# 1. 중복된 값을 제거한다.
# 2. 이상치를 대체한다.
# (변수 특성상 가장 높은 이상치는 박스플롯의 100%, 가장 낮은 이상치는 제거
# 위 두과정을 거친 뒤 결측치를 평균으로 대체한다
```

2 - 1. 데이터 확인하기 버전 2

```
tips.columns, tips.values # 데이터가 어떻게 생겼는지 한 번 확인해봅니

tips.duplicated(), tips.isnull().sum()
# 데이터의 중복값과 결측치를 확인해봅니다.

tips.max(), tips.min(), tips.mean()
# 최댓값과 최솟값이 데이터의 특성을 고려했을 때 이상치인지 확인해봅니다.
# 즉 이상치가 있는지 알아봅니다.
```

3. 전처리 시작 전에 원본 파일에서 평균 구하기 (비교용)

```
tips_original = tips
mean_before= tips.tip.mean()
```

4. 중복된 값 제거하기

```
# 중복된 값을 제거할 때 중복된 값들 중에서 가장 큰 값을 남기려고한다.
# 이름이 같다면 팁을 많이 낸 데이터를 제외하고 삭제

# 1. 정렬을 한다.
tips.sort_values(['이름','tip'], ascending=False)
# 2. 첫번째 값이외를 제거한다.
tips.drop_duplicates('이름', keep='first') # 해당 코드가 실행되는
tips.drop_duplicates('이름', keep='first', inplcae=True) # 해당
tips # 중복 값 제거가 제대로 되었는지 확인합니다.
```

4. 이상치 제거하기

```
# 박스플롯의 최댓값과 최솟값 이외에 있는 아웃라이어를 대체, 제거한다.
QR1 = tips.tip.quantile(q = 0.25) # tip 열 박스의 25% 값
QR3 = tips.tip.quantile(q = 0.75) # tip 열 박스의 75% 값

QR2 = QR3 - QR1 # 중앙값 (Q3 - Q1)
```

```
max = QR2 + (QR3 * 1.5) # tip 열 박스의 최댓값
min = QR2 - (QR1 * 1.5) # tip 열 박스의 최솟값
```

```
# 너무 높은 이상치 박스 플롯 100% 값으로 대체
filter1 = tips['tip'] > max
tips.loc[filter1, 'tip'] = max

# 너무 낮은 이상치 삭제하기
# 박스플롯의 0%는 음수가 나올 수 도 있으므로 min가 아니라 0으로 대체하거나
filter2 = tips['tip'] < min
tipd.drop(index=tips[filter2], inplace=True)
```

5. 중복, 이상치 제거 대체 후 평균을 오리지널 평균과 비교

```
mean_after = round(tips.tip.mean(),2) # 확인하기 편하게 소수점 2째
print(mean_before, mean_after)

# 두 평균을 비교한다.
# 이상치와 중복된 값이 꽤 있어 차이가 날 것으로 예상된다.
# 그러므로 결측치를 대체하기 위한 평균은 데이터 클린징한 후의 평균을 사용한
```

6. 결측치 평균으로 대체하기

```
tips.tip.fillna(mean_after.tip, inplace=True) # 결측치를 mean_a
tips.tip.dropna() # 열에 결측치가 있으면 해당 열을 포함한 행 삭제
tips.dropna()
```

전처리 끝

7. 그룹화

흡연 여부에 따른 각 요일의 평균 및 중위값 지불 금액을 계산하세요.

```
# 흡연자로 먼저 그룹핑하고 흡연자의 요일별 그룹, 비흡연자의 요일별 그룹핑을
# 그룹핑한 데이터프레임의 평균과 중위값을 집계함수 agg를 이용해 표현한다.
smoker_day_tips = tips.groupby(by=['smoker','day'])['tip'].ag
smoker_day_tips.round(2)
```

		mean	median
smoker	day		
Yes	Thur	3.03	2.56
	Fri	2.71	2.50
	Sat	2.88	2.69
	Sun	3.52	3.50
No	Thur	2.67	2.18
	Fri	2.81	3.12
	Sat	3.10	2.75
	Sun	3.17	3.02

8. 재구조화

```
pd.pivot_table(smoker_day_tips,
                index='smoker',columns='day',values='mean').r
```

day	Thur	Fri	Sat	Sun
smoker				
Yes	3.03	2.71	2.88	3.52
No	2.67	2.81	3.10	3.17

각 그룹에서 가장 많은 팁을 받은 인원의 정보 출력

```
most_tipped = tips.groupby(by=['sex', 'day']).apply(lambda x:
```

pd.cut구간 나눠서 1~6등급 만들고

merge로 병합하기

```
x = df.국어+df.영어+df.수학+df.과학+df.사회
df['과목평균'] = x/5
a = df['과목평균']
print(df)
bins = [0, 50,60,70,80,90,100]
labels=(['6등급','5등급','4등급','3등급','2등급','1등급'])

cuts = pd.cut(a, bins, right=True, labels=labels)
pd.merge(df, cuts, left_index=True, right_index=True)
```

조건에 맞게 삭제하기 → 이상치 제거 대용

```
# 키가 200 이상이면 삭제
filter = df[’키’] > 200
df.drop(index=df[filter].index
```

조건에 맞게 값 수정 → 결측치 제거 대용

```
# 키가 200 이상이면 평균으로 대체
filter = df[’키’] > 200
df.loc[filter,'키'] = df.키.mean()
```

열의 NaN 대체하기

```
df['SW특기'].fillna('초보')
df['키'].fillna(df.키.mean())
```