빅데이터 분석기사 실기 정리 문서

**Chap1. Pandas Practice**

▶ Check DataFrame Information

1. df.info()

- 각 Column의 Data type ( dtypes ) 정보.

- Null 값의 개수

- DataFrame Memory usage.

|  |
| --- |
| <class 'pandas.core.frame.DataFrame'>  Index: 3500 entries, 0 to 3499  Data columns (total 9 columns):  # Column Non-Null Count Dtype  --- ------ -------------- -----  0 총구매액 3500 non-null int64  1 최대구매액 3500 non-null int64  2 환불금액 1205 non-null float64  3 주구매상품 3500 non-null object  4 주구매지점 3500 non-null object  5 내점일수 3500 non-null int64  6 내점당구매건수 3500 non-null float64  7 주말방문비율 3500 non-null float64  8 구매주기 3500 non-null int64  dtypes: float64(3), int64(4), object(2)  memory usage: 273.4+ KB |

Table 1. df.info() 결과

2. df.describe()

- 수치형 데이터에 대한 요약 통계 제공 ( mean, std, min, quantile, max )

- 옵션(include=’object’)을 통해 문자열(object) 자료도 열람 가능.

|  |
| --- |
| 총구매액 최대구매액 ... 주말방문비율 구매주기  count 3.500000e+03 3.500000e+03 ... 3500.000000 3500.000000  mean 9.191925e+07 1.966424e+07 ... 0.307246 20.958286  std 1.635065e+08 3.199235e+07 ... 0.289752 24.748682  min -5.242152e+07 -2.992000e+06 ... 0.000000 0.000000  25% 4.747050e+06 2.875000e+06 ... 0.027291 4.000000  50% 2.822270e+07 9.837000e+06 ... 0.256410 13.000000  75% 1.065079e+08 2.296250e+07 ... 0.448980 28.000000  max 2.323180e+09 7.066290e+08 ... 1.000000 166.000000  [8 rows x 7 columns] |

Table 2. df.describe() 결과

|  |
| --- |
| 주구매상품 주구매지점  count 3500 3500 # 각 Column별 Row 개수  unique 42 24 # count 개수의 Row 중 고유한 값의 도메인 개수  top 기타 본 점 # 전체에서 가장 많이 존재하는 값  freq 595 1077 # top 값의 빈도(출현) 수 |

Table 3. df.describe( include=’object’ ) 결과

3. df.shape

- DataFrame의 행과 열을 튜플 형태로 확인. (ex. (20, 5) 🡪 20행 5열. (13, 10) 🡪 13행 10열. )

4. df.head(N)

- 첫 Row를 포함한 N개의 Row만 출력 (N 입력이 없으면 기본값 5 )

|  |
| --- |
| 총구매액 최대구매액 환불금액 주구매상품 ... 내점일수 내점당구매건수 주말방문비율 구매주기  cust\_id ...  0 68282840 11264000 6860000.0 기타 ... 19 3.894737 0.527027 17  1 2136000 2136000 300000.0 스포츠 ... 2 1.500000 0.000000 1  2 3197000 1639000 NaN 남성 캐주얼 ... 2 2.000000 0.000000 1  3 16077620 4935000 NaN 기타 ... 18 2.444444 0.318182 16  4 29050000 24000000 NaN 보석 ... 2 1.500000 0.000000 85  [5 rows x 9 columns] |

Table 4. df.head() 또는 df.head(5)의 결과

5. df.tail(N)

- 마지막 Row를 포함한 하위 5개의 Row만 출력 (N 입력이 없으면 기본값 5)