결측치 처리

조사연월: 202400으로 다른 행들과 동일하게 채워넣어줌.

가구구분코드: 1과 2로 나누어져 있음.

데이터셋이 정렬되어 있기 때문에 위의 행의 값으로 텍스트, 스크린샷, 폰트, 영수증이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.채워넣음.(ffill)

가구원수: 가구구분코드와 마찬가지로 그룹화 되어 있기 때문에 위의 행의 값으로 채워넣음. (ffill)

취업인원수: ffill

가구주\_연령: ffill

노인가구여부: 가구주\_연령이 65세 이상이면 1로

텍스트, 영수증, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

가구주\_학력코드: 8이 있으면 NaN값으로 바꾼 후 ffill

텍스트, 번호, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

가구주\_취업여부: ffill

텍스트, 영수증, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

가구주\_10차산업분류코드: ffill

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

거처구분코드: ffill

텍스트, 번호, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

입주형태코드: ffill

텍스트, 번호, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

전국전체가구\_1인이상소득10분위코드: 소득구간코드로 예측

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

소득구간코드: 전국전체가구\_1인이상소득10분위코드로 예측

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

소득구간코드 있고, 분위코드 없음: 같은 소득구간에 속한 사람들의 평균 분위값 계산

분위코드 있고, 소득구간 코드 없음: 분위코드별 평균 소득구간 계산

둘다 없으면: 가구원수를 보고, 같은 가구원수를 가진 사람들의 평균 소득구간으로 소득구간을 채움.

가계지출금액: 소비지출 + 비소비지출(세금, 보험…)

가계지출\_소비지출금액에 보정계수 1.1곱해서 계산

가계지출금액 = 가계지출\_소비지출금액 \* 1.1

가계지출\_소비지출금액, 12개 소비 항목: 순수 생활 소비 지출

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

도시읍면부구분코드: 최빈값으로 채움. 도시가 훨씬 많기 때문.

텍스트, 영수증, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

EDA

도표, 스크린샷, 라인, 그래프이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

도표, 스크린샷, 라인, 그래프이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

Outlier가 많은 이유: 소비 항목은 특성상 원래 편향된 분포를 가짐. (Right-skewed)

Outlier가 너무 많기 때문에 Outlier를 전부 clip해주는 건 특성 반영을 제대로 하지 못할 것 같아서, 극단적으로 너무 큰 값들만 clip 해줘야 할 것 같음.

이번엔 1, 3, 5인 가구에 대해 histogram으로 소비 항목별 지출 histogram plot

도표, 스크린샷, 그래프, 라인이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

스크린샷, 도표, 라인, 그래프이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

스크린샷, 도표, 그래프, 라인이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

스크린샷, 도표, 그래프, 라인이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

스크린샷, 도표, 그래프, 라인이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

스크린샷, 그래프, 도표, 라인이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

대부분이 right-skewed인걸 확인할 수 있음.

Outlier\_df.csv에 각 가구원수의 각 항목별로 몇개의 outlier가 있는지 저장해놨음.

Feature creation

각 품목에 log값 취해준 열들, 품목 소비 / 소비지출 해서 비율 구해준 열들 dataset에 붙여서 ratio.csv로 저장.

Encoding

가구구분코드: one-hot encoding

텍스트, 스크린샷, 폰트, 영수증이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

노인가구여부: 이미 one-hot encoding 되어있음.

가구주\_학력코드: 이미 ordinal encoding 되어있음.

텍스트, 번호, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

가구주\_취업여부: one-hot encoding 해야함.

텍스트, 영수증, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

가구주\_10차산업분류코드: one-hot encoding 해야함. 하지만 one-hot encoding을 하면 차원 수가 너무 많아지기 때문에 그룹화해서 one-hot encoding 진행.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

거처구분코드: one-hot encoding 해야함.

텍스트, 번호, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

입주형태코드: one-hot encoding 해야함.

텍스트, 번호, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

전국전체가구\_1인이상소득10분위코드: 이미 ordinal encoding 되어있음.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

소득구간코드: 이미 ordinal encoding 되어있음.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

도시읍면부구분코드: 이미 one-hot encoding 되어있음.

텍스트, 영수증, 스크린샷, 폰트이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

Feature engineering

Modeling