■ TPO 기반 소비패턴 분석 중간 현황_12월 1주차

	상세내용	비고
데이터 설명	1. 기반 데이터 - 2018. 7~9월 3개월 매출 건 중 1건이라도 승인이 있던 고객의 전국 매출 데이터 ※ 1년 매출이 아닌 3개월 매출로서, 데이터에 계절성이 나타날 수 있음(예: 휴가철, 환절기, 이사철) - 매출 정보: 승인금액, 날짜, 시간, 가맹점, 업종 - 고객 정보: 나이, 성별, 집주소, 직장주소 / 가맹점 정보: 행정동 주소, 업종명 - 매출건수: 약 39억 건, 고객 수: 약 487만 명, 가맹점 수: 129만 개, 행정동 수: 3573개	
이슈 및 처리현황	2. 데이터 처리 - 이슈: EDW에서 seg별 주 소비지역 cut-off 임계값을 산출하는 과정에서 지속적인 failure 발생 - 조치사항: 전체 고객 모집단과 demo 비율이 유사한 stratified sample (n=50만명)을 이용하여 임계값을 산출하기로 변경 ※ 본래 목적은 서울 일부 지역의 평균값을 이용한 주 소비지역(A)과 전국 전체 지역의 평균값을 이용한 주소비지역(B) 을 비교하는 것이었으나, 모수와 최대한 동일하게 구성한 샘플에서 추출한 값(B')을 사용하더라도 큰 문제 없을 것으로 판단 ※ (만약, 샘플의 대표성 문제로 실제 모수의 값과 크게 다르다고 판단될 경우, 샘플 수에 따른 값의 변화를 시뮬레이션하여 전체 모수의 평균치 추출 가능) 3. 현황 - 현재까지의 산출물, 접근방식, 코드 및 알고리즘 설명 완료 - 정제된 데이터, 소스코드, 결과물, 백업파일은 ./src/S05214 폴더에 저장되어 있음	1/2

■ TPO 기반 소비패턴 분석 중간 현황_12월 1주차

	상세내용	비고
데이터 설명	1. 기반 데이터 - 2018. 7~9월 3개월 매출 건 중 1건이라도 승인이 있던 고객의 전국 매출 데이터 ※ 1년 매출이 아닌 3개월 매출로서, 데이터에 계절성이 나타날 수 있음(예: 휴가철, 환절기, 이사철) - 매출 정보: 승인금액, 날짜, 시간, 가맹점, 업종 - 고객 정보: 나이, 성별, 집주소, 직장주소 / 가맹점 정보: 행정동 주소, 업종명 - 매출건수: 약 39억 건, 고객 수: 약 487만 명, 가맹점 수: 129만 개, 행정동 수: 3573개	
이슈 및 처리현황	2. 데이터 처리 - 이슈: EDW에서 seg별 주 소비지역 cut-off 임계값을 산출하는 과정에서 지속적인 failure 발생 - 조치사항: 전체 고객 모집단과 demo 비율이 유사한 stratified sample (n=50만명)을 이용하여 임계값을 산출하기로 변경 ※ 본래 목적은 서울 일부 지역의 평균값을 이용한 주 소비지역(A)과 전국 전체 지역의 평균값을 이용한 주소비지역(B) 을 비교하는 것이었으나, 모수와 최대한 동일하게 구성한 샘플에서 추출한 값(B')을 사용하더라도 큰 문제 없을 것으로 판단 ※ (만약, 샘플의 대표성 문제로 실제 모수의 값과 크게 다르다고 판단될 경우, 샘플 수에 따른 값의 변화를 시뮬레이션하여 전체 모수의 평균치 추출 가능) 3. 현황 - 현재까지의 산출물, 접근방식, 코드 및 알고리즘 설명 완료 - 정제된 데이터, 소스코드, 결과물, 백업파일은 ./src/S05214 폴더에 저장되어 있음	1/2

■ TPO 기반 소비패턴 분석 중간 현황_12월 1주차

	상세내용	비고
전체 매출 분석 현황 요약	1. 주 소비지역 추출 - Pandas 라이브러리의 DataFrame 으로 작업 - 작업 방식 1) 고객번호, 주 소비요일, 주 소비시간, 행정동 을 기준으로 그룹핑 2) 그룹핑된 값에 고객 ID를 데모 정보와 join 3) 고객 데모 seg별 cut-off criteria 를 넘기는 매출 건만 추출 → "주 소비지역" 4) 추출된 주 소비지역을 행정동 코드와 join 5) 추출된 주 가맹점 ID를 업종 정보와 join 2. 아웃풋 - 고객별 주 소비지역 임계값 매칭 완료 (소스코드): tpo-cutoff-criteria.ipynb - 고객별 주 소비지역 그룹핑 완료 (소스코드): tpo-groupby-cutoff.ipynb - 고객별 주 소비지역 그룹핑 결과물: trx_tpo_gb01.csv - 고객별 주 소비지역 행정동 및 데모정보 조인 (소스코드): tpo-groupby-join.ipynb 3. 확인사항 - Spark, Impala, Hive 등 대용량·분산처리 툴 이용 가능시점 - Rstudio 등 분석 툴 이용 가능시점 - 세부 과제 정의	전국 지역 50만 명 대상 cut-off 임계값 필요

2/2

■ TPO 기반 소비패턴 분석 중간 현황_12월 1주차

	상세내용	비고
전체 매출 분석 현황 요약	1. 주 소비지역 추출 - Pandas 라이브러리의 DataFrame 으로 작업 - 작업 방식 1) 고객번호, 주 소비요일, 주 소비시간, 행정동 을 기준으로 그룹핑 2) 그룹핑된 값에 고객 ID를 데모 정보와 join 3) 고객 데모 seg별 cut-off criteria 를 넘기는 매출 건만 추출 → "주 소비지역" 4) 추출된 주 소비지역을 행정동 코드와 join 5) 추출된 주 가맹점 ID를 업종 정보와 join 2. 아웃풋 - 고객별 주 소비지역 임계값 매칭 완료 (소스코드): tpo-cutoff-criteria.ipynb - 고객별 주 소비지역 그룹핑 완료 (소스코드): tpo-groupby-cutoff.ipynb - 고객별 주 소비지역 그룹핑 결과물: trx_tpo_gb01.csv - 고객별 주 소비지역 행정동 및 데모정보 조인 (소스코드): tpo-groupby-join.ipynb 3. 확인사항 - Spark, Impala, Hive 등 대용량·분산처리 툴 이용 가능시점 - Rstudio 등 분석 툴 이용 가능시점 - 세부 과제 정의	전국 지역 50만 명 대상 cut-off 임계값 필요

2/2