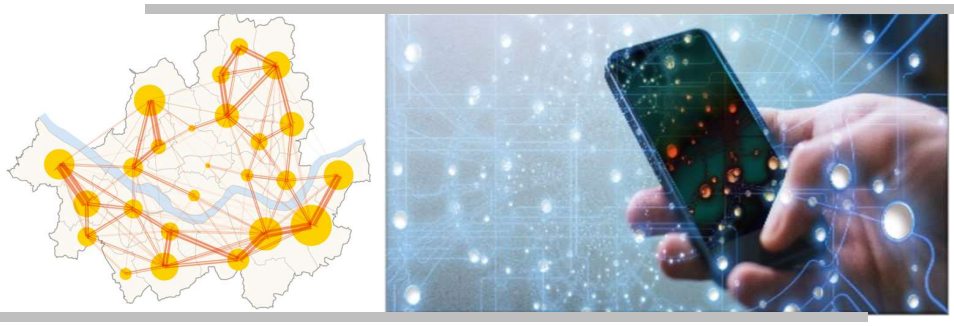
2022년

서울 시민생활 데이터 매뉴얼



서울특별시

목 차

1. 서울 시민생활 데이터 개요

❶ 개발 배경 및 필요성

- 증가하는 1인가구의 특성을 파악할 수 있는 데이터 확보 필요성 대두
- 공공・민간데이터 가명결합을 통해 1인가구의 삶의 특성을 파악하고, 삶의 질을 향상시킬 수 있는 정책 활용을 위한 데이터 개발 및 분석을 목적으로 함

데이터 가명 결합

공공 행정 분야, 민/관 가명결합 데이터 첫 활용 사례

- 급격한 인구구조 변화, 증가하는 1인가구의 삶의 질에 대한 데이터 필요성 대두

↓

- 데이터 확보: 통계청-통신사 데이터 가명 결합

↓

- 분석: 세분화된 1인가구 집단의 수, 삶의 질과 관련한 지표 개발

↓

- 월 단위 지속적 데이터 확보
  * 기존 자치구/년 단위 ⇒ 행정동/월 단위 생산 활용

1인가구의 삶의 질 이해를 위한 지표 개발

사회적관계, 주말여행, 양극화 등 1인가구의 삶 조명

- Insight1: 다양한 1인가구의 삶
  * 1인가구 내에서도 소득, 여가 등에서 큰 차이가 존재함, 40~50대 1인가구의 재정위기 등

↓

- Insight2: 성/연령/지역 등에 따른 1인가구 수 차이
  * 곳곳에 숨어있는 40대 1인가구, 20대와 함께 사는 20대 1인가구 등 확인

↘

- Insight3: 여가, 사회적 관계, 재정 상태 등 새로운 지표
  * 통화/SNS 등으로 1인가구의 사회관계성을 진단하고, 휴대폰 요금 연체 등으로 재정 위기 등 진단

↓

- 자치구/행정동 별 차별화된 1인가구 정책 수립
  * 사회관계, 이동이 적은 노년층 밀집지역: 스마트 헬스케어, 외출 지원 사업 등
  * 서울시 청년정책 및 1인가구 지원 사업 등에 활용

❷ 데이터 개요

- 1인 가구를 행정동과 성/연령으로 세분화
- 결합데이터를 이용한 1인가구 집단의 세분화

	1인 가구 수	커뮤니케이션: 커뮤니케이션이 적은 집단	이동 (외출/여행): 외출이 적은 집단, 출근소요시간 및 근무시간이 많은 집단	여가 (영상서비스 소비): 동영상서비스 이용이 많은 집단	생활 (생활서비스): 생활서비스 이용이 많은 집단	재정 (재정위기): 재정 상태에 대한 관심 집단
항목	행정동 > 성/연령대 (월 별, 20세 이상, 5세 단위)	통화/SNS 이용이 적은 1인 가구수	1인가구의 평일/휴일 이동량	1인가구의 영상 서비스 소비량	1인가구의 생활서비스 이용량	재정 위기상황인 1인 가구수
의미	결합데이터를 이용하여 추정한 행정동, 성/연령별 1인 가구 수	온라인 커뮤니케이션이 적은 1인 가구	평일/휴일 거주지 밖으로 외출이 적은 1인 가구, 출근시간이 길고 직장 인근 체류시간도 긴 집단	여가 관련, 유튜브/넷플릭스 등 영상 서비스 소비가 많은 1인 가구 구분	생활 편의 영위 정도, 모바일 의존도 파악, 금융/쇼핑/배달/게임 서비스 이용 형태	지원 필요 수준의 재정위기가 예상되는 1인 가구 구분
활용 정보	• 휴대폰 이용정보 • 모바일 서비스 이용 패턴	• 통화/문자/SNS 대상자 수, 건 수 등	• 평일/휴일 외출 건 수, 거리 등 • 출근 소요시간 • 직장 인근 체류 시간	• 유튜브 등 동영상 서비스 데이터 이용량	• 모바일 서비스 이용량 (사용 일 수)	• 소액결제 • 휴대폰 요금 연체 여부
모형 개발	• 1인가구 확률추정 및 지수 추정을 위한 Logistic Regression 모형 개발: 각 개인별 0~1점 사이의 예측 점수 제공 (1인가구일 확률, 지수 도출을 위한 스코어 등)					

3 용어 정리

유형	용어	정의 및 산출방법	단위
커뮤니케이션	월평균 통화량	• 최근 3개월[1] 간 월 평균 통화 건 수 • 원천자료: 음성통화 수발신 건 수 • 산출방법[2]: 최근 3개월 총 통화 건 수/3개월	건수
	월평균 문자량	• 최근 3개월간 월 평균 문자 건 수 • 원천자료: 문자 수발신 건 수 • 산출방법: 최근 3개월 총 문자 건 수/3개월	건수
	월평균 통화대상자 수	• 최근 3개월의 월 평균 통화 대상자 수 • 원천자료: 음성통화 수발신 대상 전화번호 • 산출방법: 최근 3개월 총 통화 대상자 수 /3개월	번호수
	월평균 문자대상자 수	• 최근 3개월의 월 평균 문자 대상자 수 • 원천자료: 문자 수발신 대상 전화번호 • 산출방법: 최근 3개월 총 문자 대상자 수/3개월	번호수
	월평균 SNS 사용횟수	• 최근 3개월의 월 평균 SNS 사용지수 • 원천자료: 최근 3개월 SNS 사용 횟수 • 산출방법: 원천자료를 Z-score[3]로 변환	Z-Value
재정	월평균 소액결제 사용횟수	• 최근 3개월 월 평균 소액결제 사용 횟수 • 원천자료: 소액결제 사용횟수 • 산출방법: 최근 3개월 총 소액결제 사용 횟수/ 3개월	횟수
	월평균 소액결제 사용금액	• 최근 3개월의 월 평균 소액결제 사용 금액 • 원천자료: 소액결제 사용금액 • 산출방법: 최근 3개월 총 소액결제 금액 / 3개월	원
	최근 3개월 내 요금 연체 비율	• 최근 3개월 내 요금 연체 비율 • 원천자료: 요금연체 유무 • 산출방법: 전체 인구대비 요금연체 인구 비율	%

1) "최근 3개월"은 모두 데이터 집계시점을 기준으로 이전 3개월을 의미함. 예를 들어 데이터가 6월에 집계되었다면 이전 3개월은 3-5월을 의미함 ※ 6월 데이터는 최근 3개월(3~5월) 기준으로 산출, 익월(7월) 공개

2) 산출방법은 한 명을 기준으로 값을 산출하는 방법을 나타냄. 예를 들어 어떤 사람의 월평균 통화량은 3개월간 총통화 건수를 3으로 나눈 값임. 행정동별, 성별, 연령별 요약통계량은 이렇게 계산된 개인별 월평균 통화량을 이용하여 집계됨.

3) 계산된 값을 전체 집단에 대해서 표준화하여 제공함. 행정동별, 성/연령별 상대적인 비교는 가능하나 집계시점의 수집된 데이터의 전체 평균과 분산으로 표준화되어 시계열 비교로 적합하지 않음.

유형	용어	개념	단위
이동 (평일/ 휴일)	야간상주지 변경 횟수	• 최근 36개월 기준 거주지 추정 위치[1]의 변경 횟수 (상주지는 매월 추정)	횟수
	주간상주지 변경 횟수	• 최근 36개월 기준 근무지 추정 위치[2]의 변경 횟수 (상주지는 매월 추정)	횟수
	평일 총 이동거리 합계	• 최근 3개월 월 평균 평일 총 이동거리 • 산출방법: 최근 3개월 평일 총 이동거리/3개월 평일의 일수	Km/1개월 평일
	휴일 총 이동거리 합계	• 최근 3개월 월 평균 휴일 총 이동거리 • 집계 방식: 최근 3개월 내 휴일 총 이동거리/3개월 휴일의 일수	Km/1개월 휴일
	평일 총 이동 횟수	• 최근 3개월의 월 평균 평일 총 이동 횟수[3] • 산출방법: 최근 3개월 내 평일 총 이동 횟수/3개월 평일의 일수	횟수/1개월 평일
	휴일 총 이동 횟수	• 최근 3개월의 월 평균 휴일 총 이동 횟수[3] • 산출방법: 최근 3개월 휴일 총 이동 횟수/3개월 휴일	횟수/1개월 휴일
	집추정 위치 평일 총 체류시간	• 최근 3개월의 월 평균 평일 총 야간상주지 근처 체류시간 • 산출방법: 최근 3개월 평일 총 야간상주지 근처 체류시간/3개월 평일	분/1개월 평일
	집추정 위치 휴일 총 체류시간	• 최근 개월의 월 평균 휴일 총 야간상주지 근처 체류시간 • 산출방법: 최근 3개월 휴일 총 야간상주지 근처 체류시간 / 3개월 휴일	분/1개월 휴일

1) 야간상주지로 01시~06시 휴대폰 위치 기준 1개월 동안 50% 이상의 시간을 체류한 행정동
2) 주간상주지로 11시~15시 휴대폰 위치 기준 1개월 동안 50% 이상의 시간을 체류한 행정동
3) 이동은 야간상주지(집 추정 위치) 추정 행정동에서 다른 행정동으로의 이동을 의미하며 하루동안 다수의 이동이 관측될 수 있음

유형	용어	개념	단위
기타 이동	평균 출근 소요시간	• 주중 출근 편도 소요시간을 기준으로 계산되며 1개월간 해당 소요시간의 평균값을 의미 • 180분 이상 값은 절단하여 제공 • 산출방법: 최근 1개월 평일 총 출근 소요시간/평일 수	분/출근 주중평균
	평균 근무시간	• 주중 근무시간 기준 최근 1개월 평균 • 710분 이상 값은 절단하여 평균 계산 • 산출방법: 최근 1개월 평일 총 근무시간/평일 수	분/평일 주중평균
	월평균 지하철 이동일수	• 최근 3개월의 월 평균 지하철 이용일수 • 산출방법: 최근 3개월 총 지하철 이용일수 / 3개월	일
영상 서비스	월평균 동영상/방송 사용량	• 최근 3개월의 월 평균 동영상/방송 서비스 사용일수 • 사용량은 각 서비스 이용 일수의 합[1]으로 정의 • 산출방법: 최근 3개월 총 동영상/방송 서비스 사용일수의 합/ 3개월	일
	월평균 유튜브 사용일수	• 최근 3개월의 월 평균 유튜브 사용일수 • 산출방법: 최근 3개월 총 유튜브 사용일수/3개월 • 해당 날짜의 사용 여부는 서비스의 접속 여부로 계산하며, 원천자료를 Z-score로 변환하여 제공	Z-Value
	월평균 넷플릭스 사용일수	• 최근 3개월의 월 평균 넷플릭스 사용일수 • 산출방법: 최근 3개월 총 넷플릭스 사용일수/ 3개월 • 해당 날짜의 사용 여부는 서비스의 접속 여부로 계산하며 원천자료를 Z-score로 변환하여 제공	Z-Value

1) 하루 동안 동영상 범주의 서비스와 방송 범주의 서비스를 동시에 사용했다면 각각의 사용량을 계산. 값은 개별 콘텐츠 서비스 단위로 산정되어 이용한 서비스의 총 수가 해당 값으로 산정됨
ex) 동영상 관련 3개의 서비스 이용 + 방송 관련 하나의 서비스 이용 시 산정된 갓은 4

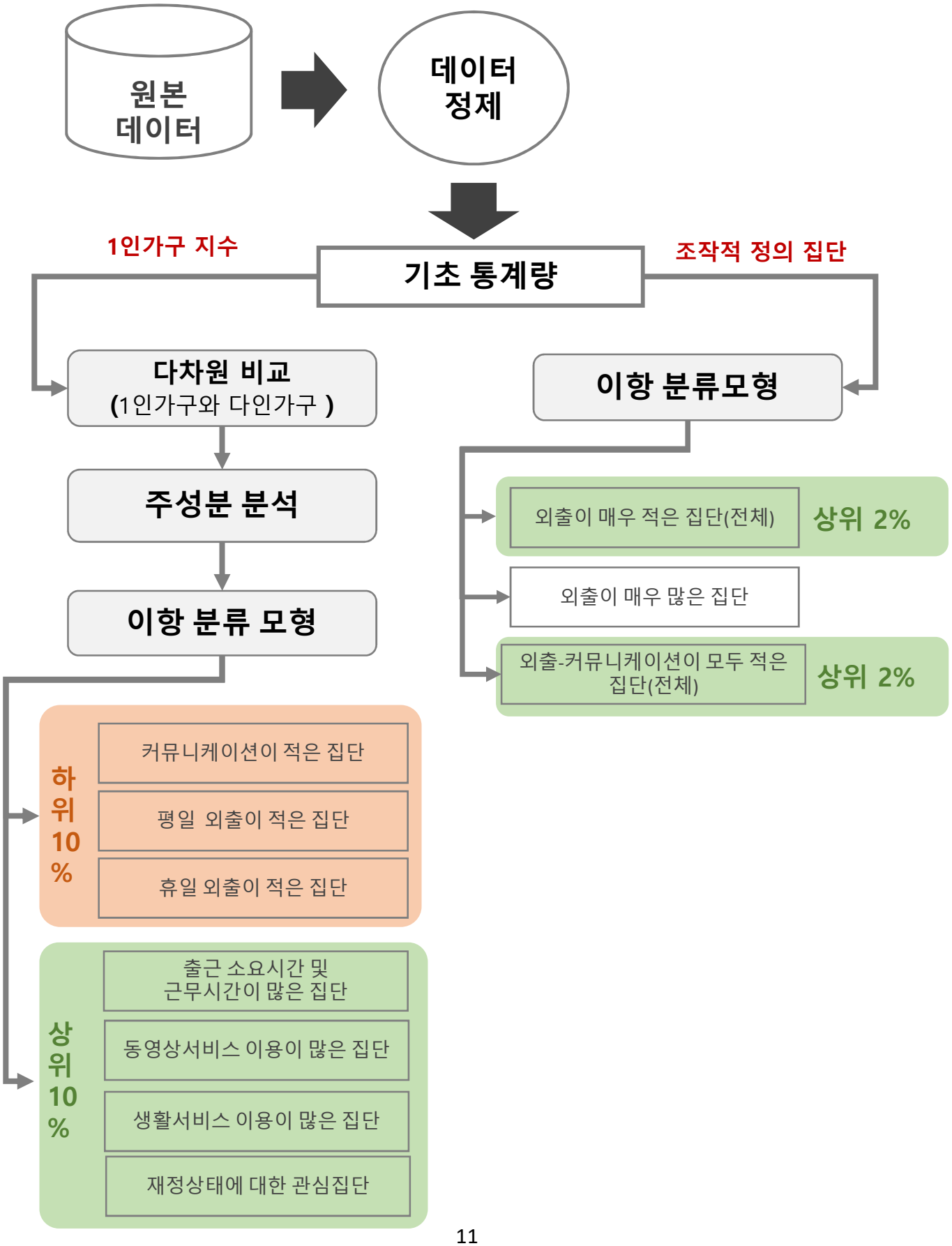
유형	용어	개념	단위
생활서비스 서비스	월평균 게임 사용량	• 최근 3개월의 월 평균 게임 서비스 사용량 • 사용량은 각 서비스 이용일수의 합으로 정의 • 집계 방식: 최근 3개월 총 게임 서비스 사용일수의 합 / 3개월	일
	월평균 금융 사용량	• 최근 3개월의 월 평균 금융 서비스 사용일수 • 사용량은 각 서비스 이용일수의 합으로 정의 • 집계 방식: 최근 3개월 총 금융 서비스 사용일수의 합 / 3개월	일
	월평균 쇼핑 사용량	• 최근 3개월의 월 평균 쇼핑 서비스 사용일수 • 사용량은 각 서비스 이용일수의 합으로 정의 • 집계 방식: 최근 3개월 총 쇼핑 서비스 사용일수의 합 / 3개월	일
	월평균 배달 사용일수	• 최근 3개월의 월 평균 배달 서비스 사용일수 • 집계 방식: 최근 3개월 총 배달 서비스 사용일수 / 3개월 • 해당 날짜의 사용 여부는 서비스의 접속 여부로 판단	일
	월평균 배달 브랜드 사용일수	• 최근 3개월의 월 평균 배달 브랜드 서비스 사용일수 • 집계 방식: 최근 3개월 총 배달 브랜드 서비스 사용일수 / 3개월 • 해당 날짜의 사용 여부는 서비스 접속 여부로 판단	일
	월평균 배달 식재료 사용일수	• 최근 3개월의 월 평균 배달 식재료 서비스 사용일수 • 집계 방식: 최근 3개월 총 배달 식재료 서비스 사용일수 / 3개월 • 해당 날짜의 사용 여부는 서비스 접속 여부로 판단	일
-	월평균 데이터 사용량	• 최근 3개월의 월 평균 데이터 이용량 • 집계 방식: 최근 3개월 총 데이터 이용량 / 3개월	GB

2. 분석 절차 및 방법

❶ 데이터 결합 과정과 방식

기획 단계	서울시, 통계청, SKT	• 1인가구 정책 수립 관련 서울시 정책부서 요구 분석 • 통계청/SKT 데이터 이용, 분석 가능한 방향/내용 기획
결합 데이터 수집	통계청	• 등록센서스, 경제/사회 통계 등 내부 데이터 수집 • 결합 key 항목 확인: 생년월일, 이름 등
	SKT	• 가입정보, 요금/단말기, 기호/관심사, 위치정보 등 수집 • 결합 key 항목 확인: 휴대폰, 생년월일, 이름 등
협업 계약	서울시, 시립대, 통계청, SKT	• 1인가구 분석 내용, 활용 데이터 및 일정 확정 • MOU: 사업 추진 확정
1차 반출심사	통계청, SKT	• 분석 기획 내용을 감안해 데이터 항목/기간, 대상자 등 가명결합에 필요한 데이터 확정: 통계청/SKT 내부 심사
데이터 추출/가공	통계청, SKT	• 1차 심사결과 준수, 필요 데이터 추출, 개인식별이 불가능하도록 데이터 구간화, z-score 변환 등 비식별 처리
2차 반출 심사	통계청, SKT	• 1차 심사결과 준수 여부 및 개인식별 가능성 검증 • 결합기관 전송 의사 결정
결합 기관 전송	통계청, SKT	• 결합 기관은 통계청(개인정보보호위원회 승인 기관) • 통계청, SKT에서 결합기관에 데이터 전송
결합 기관 반출 심사	통계청	• 가명결합에 대한 법/제도적 절차에 따라 외부 심사위원 대상, 항목 및 재식별 위험성에 대한 적정성 검증 실시
반출 및 분석	통계청 (환경 제공)	• 통계청에서 제공한 데이터 분석센터에서 시립대/SKT가 방문, 분석 실시: 통계처리 된 결과 반출(통계청 검수)

❷ 시민생활 유형별 1인가구 추정절차



❸ 작성과정-생활유형별 1인가구수 추정

[1단계] 1인가구와 다인가구의 단순 비교

- 가구를 연령대, 성별을 기준으로 하여 작은 집단으로 구분
- 각 집단별로 결합 데이터 각 변수의 기초통계량 비교
- 기초통계량 비교를 통한 집단별 특징 도출
- **목표: 연령대와 성별에 따른 1인가구와 다인가구의 생활특성 차이 비교**

- **예시 질문**: 1인가구인 20대와 다인가구를 구성하는 20대(청년층 초기)는 통화량에 차이가 있을까요?
- **예시 답변**: [1단계]의 결과를 보면, 1인가구 20대의 경우 평균 통화량이 347.8건/월으로 다인가구 20대의 평균 통화량보다 308.4건/월 많습니다.

기초통계량 예시

연령대	가구형태	평균 통화대상자수 (명/월)	평균 통화량 (건 수/월)	평균 문자대상자 수 (명/월)	평균 문자량 (건 수/월)	SNS 사용횟수 (횟수/월)
청년층 초기 (20~29세)	1인가구	38.4	**347.8**	16.2	20.7	4823
	다인가구	33.2	**308.4**	14.1	20	4562
청년층 후기 (30~39세)	1인가구	46	344.4	18.1	25.7	4376
	다인가구	47.4	322.9	18.2	25.8	4118
중년층 (40~49세)	1인가구	47.3	336.8	17.7	29.9	3615
	다인가구	53.3	355.5	19.8	37.2	3440
장년층 (50~64세)	1인가구	45.7	331.7	14.9	20.3	3552
	다인가구	50.2	324.3	16.5	21.8	3102
노년층 초기 (65~74세)	1인가구	33.2	292.5	10	7.9	2698
	다인가구	34.1	232.3	10.6	8.1	2279
노년층 후기 (75세 이상)	1인가구	23.8	211.7	6.2	2.7	1555
	다인가구	22.6	157.2	6.7	3.5	1354

[2단계] 1인가구와 다인가구의 다차원 비교

- **분류모형을 이용하여 1인가구와 다인가구의 분류에 사용되는 주요 변수를 추출**
- **분석방법**
  1. 연령별로 1인가구와 다인가구를 분류하는 "로지스틱 회귀모형"을 적합
  2. 분류모형에 사용한 설명변수로 6-9쪽에 제시된 유형별 라이프스타일 변수를 사용 함. (예: 커뮤니케이션 유형에 해당하는 변수만을 사용하여 로지스틱 회귀모형을 적합)
  3. 적합모형에서 로지스틱 회귀계수의 크기와 [1단계] 단순비교 결과와의 정합성 검토
- **목표: 연령대와 성별에 따른 1인가구와 다인가구의 라이프스타일의 주요 차이 분석**

예시 질문: 1인가구인 30대(청년층 후기)와 다인가구를 구성하는 30대는 커뮤니케이션의 측면에서 어떤 변수를 통해 그 차이가 가장 잘 나타날까? (커뮤니케이션 관련 변수는 평균 통화대상자 수, 평균 문자대상자 수, 평균 통화량, 평균 문자량, SNS 사용횟수 등이 있다.)

예시 답변: 30대의 경우 1인가구와 다인가구를 구분하는 가장 중요한 커뮤니케이션 관련 변수는 평균 통화대상자 수(-0.17로 회귀계수의 절댓값이 가장 큼)로 나타났으며, 1인가구가 다인가구에 비해 평균 통화대상자 수가 적게 나타나는 것으로 확인되었다(평균 통화대상자 수의 회귀계수가 음수이므로 평균 통화대상자 수가 적을수록 1인가구일 가능성이 높음).

이항 로지스틱 회귀계수 예시

연령 구분	평균 통화대상자 수	평균 통화량	평균 문자량	평균 문자대상자 수	SNS 사용횟수
청년층 초기	-0.01	0.05	-0.17	0.27	0.04
청년층 후기	**-0.17**	0.11	0.01	0.02	0.13
중년층	-0.2	0.05	-0.03	-0.08	0.15
장년층	-0.21	0.1	0.1	-0.18	0.26
노년층 초기	-0.18	0.26	0.06	-0.14	0.21
노년층 후기	0.01	0.22	0	-0.2	0.08

[3단계] 1인가구 내 특징 추출

- 결합데이터 항목별(커뮤니케이션, 이동, 영상 서비스, 생활서비스, 재정상태 등) 지수 개발
    - (예시) **커뮤니케이션 관련 변수**는 평균 통화대상자 수, 평균 문자대상자 수, 평균 통화량, 평균 문자량, SNS 사용횟수로 총 5개의 변수로 구성되어 있다.
- 해당 5개 변수들에 적당한 가중치를 주어서 커뮤니케이션의 양에 대한 지표를 구성
- 1인가구의 전체 데이터에 대해서 항목별 변수들에 대해 주성분 분석을 수행하고, 계산된 제1주성분을 해당 항목의 1인가구 지수로 사용
- **목표: 커뮤니케이션, 재정상태 등 제공데이터의 유형에 따라 1인가구 지수를 계산하여 행정동, 연령, 성별 대표값을 산출**

커뮤니케이션 지수 계산식

구분	평균 통화대상자 수 (x_1)	평균 통화량 (x_2)	SNS 사용횟수 (x_3)	평균 문자대상자 수 (x_4)	평균 문자량 (x_5)
주성분 계수	0.51	0.45	0.36	0.51	0.38

$$\text{커뮤니케이션 지수} = 0.51X_1 + 0.45X_2 + 0.36X_3 + 0.51X_4 + 0.38X_5$$

- 지수식을 구성하는 계수가 모두 양수로 나타났고, 다변량변수들이 모두 커뮤니케이션의 양을 나타내는 변수기 때문에, 커뮤니케이션 지수는 커뮤니케이션의 양으로 설명할 수 있음

예시 질문: 30대를 대상으로 하였을 때, "커뮤니케이션 양"이 평균적으로 가장 높은 행정동은 어디일까?

커뮤니케이션 관심집단은 통화량에 대한 요약정보로 해석할 수 있기 때문에, 커뮤니케이션 많은 집단은 해당 행정동 1인가구중 비율이 가장 작은 행정동을 찾으면 됨.

그 외 지수 계산법

- 평일 이동지수 = $-0.64X_1 + 0.64X_2 + 0.44X_3$

구분	집추정 위치 평일 총 체류시간 (x_1)	평일 총 이동 횟수 (x_2)	평일 총 이동거리 합계 (x_3)
주성분 계수	-0.64	0.64	0.44

- 휴일 이동지수 = $-0.56X_1 + 0.58X_2 + 0.59X_3$

구분	집추정 위치 휴일 총 체류시간 (x_1)	휴일 총 이동 횟수 (x_2)	휴일 총 이동거리 합계 (x_3)
주성분 계수	-0.56	0.58	0.59

- 기타 이동지수 = $0.6\,X_1 + 0.76\,X_2 + 0.25\,X_3$

구분	평균 출근소요시간 (x_1)	평균 근무시간 (x_2)	지하철이동일수 합계 (x_3)
주성분 계수	0.6	0.76	0.25

- 영상 서비스 소비지수 = $0.64X_1 + 0.59X_2 + 0.5X_3$

구분	동영상/방송 서비스 사용일수 (x_1)	유튜브 사용일수 (x_2)	넷플릭스 사용일수 (x_3)
주성분 계수	0.64	0.59	0.5

- 생활서비스 서비스 소비지수 = $0.52\,X_1 + 0.37X_2 + 0.49X_3 + 0.56X_4 + 0.19X_5 + 0.11X_6$

구분	금융 서비스 사용일수 (x_1)	게임 서비스 사용일수 (x_2)	쇼핑 서비스 사용일수 (x_3)	배달_배달 서비스 사용일수 (x_4)	배달_브랜드 서비스 사용일수 (x_5)	배달_식재료 서비스 사용일수 (x_6)
주성분 계수	0.52	0.37	0.49	0.56	0.19	0.11

- 재정상태지수 = $0.64X_1 + 0.77X_2$

구분	소액결제 사용금액 (x_1)	최근 3개월내 요금 연체 여부 (x_2)
주성분 계수	0.64	0.77

[4단계] 1인가구 세분화 및 인구수 추정

- 3단계에서 개발된 1인가구 지수를 활용하여 관심집단을 정의
- 정의한 관심집단을 통해 독특한 1인가구 수 예측모형을 생성
- **분석 방법 (1인가구 지수)**
    1. 3단계에서 계산한 지수를 기준으로 1인가구 집단을 분류한 후 로지스틱회귀모형을 적합
        - 설명변수: SKT에서 사용 가능한 설명변수 (p6-p9 참조)
        - 반응변수: 지수값을 기준으로 분류한 1인가구 (뒤쪽의 표 참조)
    2. 결합 데이터 기준 식별 가능한 사람에 대해서 적합된 로지스틱 회귀모형을 적용하여 로짓 및 조건부 확률을 계산
    3. 조건부확률을 이용한 인구수 추정모형, 로짓값을 이용한 인구수 추정
    4. 전수화 과정을 통한 계산된 추정값을 보정
- **목표: 1인가구 지수를 이용해 특별한 1인가구의 수와 비율을 행정동별로 추정**
- **세분화된 1인가구 집단을 분류하는 모형 개발**

$$\text{관심집단에 속할 확률} = f\left(\begin{array}{c}\text{커뮤니케이션 관련 변수}\\ \text{이동 관련 변수}\\ \cdots\\ \text{서비스 관련 변수}\end{array}\right)$$

- f: 로지스틱 회귀모형
- 지수(7) x 연령(6) = 총 42개 모형 개발
- 모형에 각 개인의 변수를 대입하면, 해당 개인이 관심집단에 속할 확률(0에서 1사이의 값)을 구할 수 있음
    - (예시) 어떤 20대인 개인의 결합데이터 변수들을, 20대의 커뮤니케이션이 적은 집단 모형에 대입하면 해당 개인이 커뮤니케이션이 적은 집단에 속할 확률을 구할 수 있음

• 독특한 1인가구 집단 (반응변수)

관심집단 이름	측정기준	집단 성격
커뮤니케이션이 적은 집단	• **커뮤니케이션 지수** ➢ 전화/문자 수 발신 대상자 수, 전화/문자 수 발신 건 수, SNS 사용량 기준	하위 10%
평일 외출이 적은 집단	• **평일 이동지수** ➢ 평일기준, 집 추정 위치 체류시간, 추정거주지 외부로의 외출건수 및 이동거리 기준	하위 10%
휴일 외출이 적은 집단	• **휴일 이동지수** ➢ 휴일기준, 집 추정 위치 체류시간, 추정거주지 외부로의 외출건수 및 이동거리 기준	하위 10%
출근 소요시간 및 근무시간이 많은 집단	• **기타 이동지수** ➢ 추정 거주지와 추정 근무지 간 이동 소요시간 (출근 소요 시간) 및 근무시간, 지하철 이용량 기준	상위 10%
동영상서비스 이용이 많은 집단	• **영상 서비스 소비지수** ➢ 방송/동영상 서비스 및 유튜브, 넷플릭스 이용량 기준	상위 10%
생활서비스 이용이 많은 집단	• **생활서비스 서비스 소비지수** ➢ 금융/게임/쇼핑/배달 서비스 사용량 기준	상위 10%
재정 상태에 대한 관심 집단	• **재정상태지수** ➢ 휴대폰 소액결제 사용금액, 휴대폰 요금 연체 여부 기준	상위 10%

집단구분의 이유

- 월별로 생산되는 데이터에는 1인가구 유무를 알 수 없어, 특별한 특징을 가지는 1인가구에 대한 요약통계량을 산출할 수 없음
- 결합데이터 상에서 특징을 가지는 1인가구를 정의하고 (조작적 정의) 그것을 예측하는 모형을 구성하고자 함

1) 예를 들어 커뮤니케이션이 적은 집단은 3단계에서 얻은 커뮤니케이션 지수를 이용하여 1인가구 중 하위 10%를 식별
2) 커뮤니케이션이 적은 1인가구와 그 외 집단을 구분하는 로지스틱회귀모형 적합
3) (추정방법1) 로지스틱회귀모형을 통해 임의의 한 사람이 커뮤니케이션이 적은 1인가구일 확률 추정.
4) (추정방법2) 로지스틱회귀모형의 상수항이 잘 추정되지 않는 경우에는 로짓을 이용해 값이 큰 관측치를 커뮤니케이션이 적은 1인가구로 판별

- **세분화된 1인가구수 추정방법1**
    1. 관측가능한 행정동, 성/연령별 자료를 이용 모든 사람에 대해 로지스틱 모형을 이용하여 관심 1인가구집단에 속할 확률을 계산
    2. 행정동, 성/연령별로 구해진 확률의 합계로 1인가구 수를 집계
    3. 전수화 적용
- **세분화된 1인가구수 추정방법2**
    1. 관측가능한 행정동, 성/연령별 자료를 이용 모든 사람에 대해 로지스틱 모형을 이용하여 관심 1인가구집단에 속할 확률의 로짓을 계산 (회귀식의 값)
    2. 관측치 전체에 대해 특정 비율(상위 혹은 하위 00%)를 만족하는 임계값 계산
    3. 임계값을 기준으로 관심 1인가구의 추정 후 전수화 적용
- **다른 두 가지 추정방법을 개발한 이유**
    1. 1인가구 예측 데이터의 분포가 크게 변한 경우, 추정방법1은 추정편이가 크게 발생하며 보정계수를 찾는 것이 어려움.
    2. COVID-19 유행 시기인 데이터 결합시점(2020년)과 모형적용시점 (2022년) 이동과 관련한 변수의 큰 차이가 발생함이 확인되었고, 현재 정보로는 시점간 편이를 보정하는 것에 기술적인 어려움으로 추정방법 2를 적용함
    3. 향후 모형 고도화 과정에서 추정방법 1을 이용한 인구수 추정모형을 보정할 계획임.

예시 질문:

1. 30대 중 "커뮤니케이션 지수"가 상대적으로 낮은 1인가구가 많이 거주하는 행정동은 어디일까?
2. 행정동에서 "커뮤니케이션 지수"가 상대적으로 낮은 1인가구의 비중이 높은 곳은 어디일까?

예시 답안:

1. 30대의 "커뮤니케이션이 적은 집단"이 많이 살 것으로 추정되는 곳은 강남구 역삼1동이다.
2. 그리고 행정동 별로 30대 거주인구 수를 기준으로 보았을 때 "커뮤니케이션이 적은 집단"의 비중이 상대적으로 높은 행정동은 관악구 대학동이다.

- **독특한 1인가구 집단의 해석 예시**

관심집단 이름	해석 예시
커뮤니케이션이 적은 집단	강남구 개포1동의 30대 남성의 경우 커뮤니케이션이 상대적으로 적은 1인가구의 수는 5로 추정됨
평일 외출이 적은 집단	강남구 개포1동의 30대 남성의 경우 평일 외출이 상대적으로 적은 1인가구의 수는 5로 추정됨
휴일 외출이 적은 집단	강남구 개포1동의 30대 남성의 경우 휴일 외출이 상대적으로 적은 1인가구의 수는 5로 추정됨
출근 소요시간 및 근무시간이 많은 집단	강남구 개포1동의 30대 남성의 경우 출근 시간과 근무 시간이 많고, 지하철 이용량이 상대적으로 많은 1인가구의 수는 5로 추정됨
동영상서비스 이용이 많은 집단	강남구 개포1동의 30대 남성의 경우 방송/동영상/OTT 서비스를 상대적으로 많이 이용하는 1인가구의 수는 5로 추정됨
생활서비스 이용이 많은 집단	강남구 개포1동의 30대 남성의 경우 생활 서비스를 상대적으로 많이 이용하는 1인가구의 수는 5로 추정됨
재정 상태에 대한 관심 집단	강남구 개포1동의 30대 남성의 경우 휴대폰 소액결제를 상대적으로 많이 하고 핸드폰 요금을 연체한 1인가구의 수는 5로 추정됨

* 실제 값이 아닌 해석을 위한 예시임

4 작성과정 – 조작적 정의 집단

- **분석방법(관심집단 지수)**
  1. 연령대별로 전체 관측치 데이터들을 분할
  2. 관심집단의 기준(p21 조작적정의 기준 참고)을 이용,
     각 관측치가 정의에 해당하면 관심집단(반응변수: 1),
     해당하지 않으면 그 외 집단(반응변수: 0)으로 구분
  3. 이항 분류 분석을 이용해 임의의 관측치가 관심집단에 속할 확률 추정
     (사용된 모형: 로지스틱 회귀모형)
  4. 적합된 모형을 이용해 관심집단의 각 연령대, 성별, 행정동별로 관심집단 수를 추정
- 추가적으로, 통신사의 인구보정 정보를 이용하여 계산된 추정값을 보정
- **목표: 관심집단 지수를 이용해 관심집단의 수와 비율을 행정동별로 추정**

- **조작적 정의 집단 분류 모형 개발**

$$\text{조작적 정의 집단에 속할 확률} = f\left(\begin{array}{c}\text{커뮤니케이션 관련 변수}\\ \text{이동 관련 변수}\\ \cdots\\ \text{서비스 관련 변수}\end{array}\right)$$

- f: 로지스틱 회귀모형
- 지수(3) x 연령(6) = 총 18개 모형 개발
- 모형에 각 개인의 변수를 대입하면, 해당 개인이 관심집단에 속할 확률(0에서 1사이의 값)을 구할 수 있음
  - (예시) 임의의 20대 1인의 결합데이터 변수들을, 확률추정모형에 대입하면 해당 1인이 그 집단에 속할 확률을 구할 수 있음

- **조작적 정의 기준 (반응변수)**

조작적 정의에 사용된 변수	외출이 매우 적은 집단(전체)	외출이 매우 많은 집단	외출-커뮤니케이션이 모두 적은 집단(전체)
1인가구		O	
근로소득	근로소득 없음	3천만원 초과	근로소득 없음
기초생활수급자	X	X	X
장애인 여부	X	X	X
평일과 휴일의 이동거리 합	적음(0)		적음(0)
야간 상주지 변화량	0인 사람		
평일과 휴일의 이동 건수의 합	0인 사람		0인 사람
평일과 휴일의 체류시간의 합	관측치가 속하는 연령대의 75% 분위수값보다 큰 사람		관측치가 속하는 연령대의 75% 분위수값보다 큰 사람
휴일의 이동 경향이 높은 사람		1) 휴일의 체류시간이 관측치가 속하는 연령대의 25% 분위수 값보다 작은 사람, 또는 2) 휴일의 이동 건수가 관측치가 속하는 연령대의 75% 분위수 값보다 큰 사람, 또는 3) 휴일의 이동거리가 관측치가 속하는 연령대의 75% 분위수 값보다 큰 사람	
통화대상자 수			연령별 기준으로 25% 분위수값보다 이하인 사람

- **외출이 매우 적은, 외출-커뮤니케이션이 모두 적은 집단(전체) 집계방식**

1. 각 관측치의 결합데이터 변수와 연령별 모형을 이용해 조작적 정의 집단에 속할 확률을 계산
2. 연령대별로 모든 관측치의 조작적 정의 집단에 속할 확률을 정렬
3. 연령대별로 조작적 정의 집단에 속할 확률의 상위 2% 기준을 설정
4. 원하는 행정동, 성별, 연령 기준 등의 집계 단위를 설정
5. 해당 집계 단위에서 조작적 정의 집단으로 추정되는 대상자수

= **원하는 집계 단위의 연령에 해당하는 조작적 정의 집단에 속할 확률의 상위 2% 기준보다 확률이 큰 관측치의 개수**

- **외출이 많은 집단 집계방식**

1. 원하는 행정동, 성별, 연령 기준 등의 집계 단위에 해당하는 관측치들을 선택
2. 선택한 관측치들의 결합데이터 변수와 연령별 모형을 이용해 외출이 많은 집단에 속할 확률을 계산
3. 해당 집계 단위에서 외출이 많은 집단으로 판단되는 1인가구의 수는 다음과 같이 계산:

= **원하는 집계 단위에 속하는 관측치들의 외출이 많은 집단에 속할 확률들의 총 합**

- **SKT 시장점유율을 반영하여 전수화**

행정동 > 성/연령 별 시장점유율을 이용하여 전수 추정

- **조작적 정의 집단과 해석 예시**

관심집단	측정 기준	집단 성격
외출이 매우 적은 집단	전체 시민대상 근로소득이 없고, 주중/주말 이동건수와 이동거리가 적고, 추정거주지 체류시간이 많은 사람 기준 (외출이 매우 적은 대상자를 구분하기 위함)	**상위 2%**
외출이 매우 많은 집단	1인가구 대상 근로소득이 3천만원 초과이고, 휴일의 이동건수와 이동거리가 크고, 휴일의 추정거주지 체류시간이 적은 사람 기준 (휴일 이동경향이 높은 대상자를 구분하기 위함)	•
외출-커뮤니케이션이 모두 적은 집단(전체)	전체 시민대상 근로소득이 없고, 주중/주말 이동건수와 이동거리가 적고, 추정 거주지 체류시간이 많고, 통화 대상자 수가 적은 사람 기준	**상위 2%**

* 3개의 집단에 대한 추정대상은 모두 서울 시민 전체임
* 상위 2% (혹은 10%)의 의미는 집단에 속할 가능성이 높은 상위 집단을 의미함

관심집단	해석 예시
외출이 매우 적은 집단	강남구 개포1동의 30대 남성의 경우 근로소득이 없고, 평일과 휴일의 외출이 모두 상대적으로 적은 대상자수는 5명으로 추정됨
외출이 매우 많은 집단	강남구 개포1동의 30대 남성의 경우 근로소득이 많고, 휴일의 이동량이 상대적으로 많은 1인가구의 수는 5로 추정됨
외출-커뮤니케이션이 모두 적은 집단(전체)	강남구 개포1동의 30대 남성의 경우 근로소득이 없고, 외출과 커뮤니케이션이 모두 상대적으로 적은 대상자수는 5명으로 추정됨

* 실제 값이 아닌 해석을 위한 예시임

3. 서울 시민생활 데이터 결과 공개

❶ 데이터 공개방법

1. 제공되는 통계량

- **미추정 인구수**: 해당 변수의 관측된 값이 0인 인구수
  - (예외) 미추정 인구수가 제공되지 않는 변수: 야간상주지 변경횟수, 주간상주지 변경횟수, 평균 통화건수, 평균 문자이용 건수, 평균 통화대상자 수, 평균 문자대상자 수, 평균 데이터 사용량 등
  - **평균, 25%, 50%, 75% 분위수**
  - 해당 변수의 관측된 값이 0이 아닌 인구수만을 대상으로 계산
- (예외) **인구수 비율**
  - 최근 3개월 내 요금 연체 여부 변수(이산형 변수)는 집계 단위 대비 요금의 연체 여부가 있는 인구수의 비율을 제공함

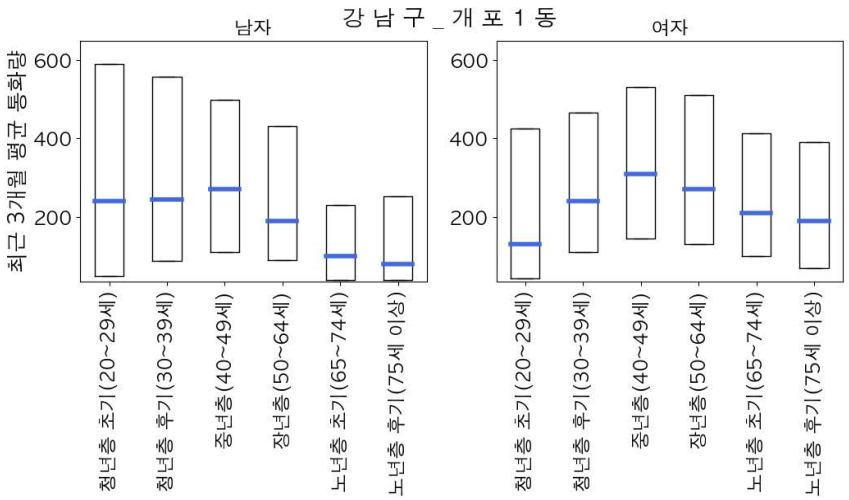
2. 마스킹(“ - ”) 처리 방식 (식별 방지)

- 미추정 인구수의 값 중 5인 이하의 값은 마스킹(masking) 처리
- 15인 이하의 추정인구로 변수들의 분위수가 계산된 경우에는 마스킹 처리

➢ **제공되는 데이터형식으로 상자그림을 그릴 수 있음**

Box-plot

* 실제 값이 아닌 예시임



➢ 제공되는 데이터 통계량(평균) 해석 예시

* 실제 값이 아닌 해석을 위한 예시임

변수	해석 예시
평균 통화량	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 총 통화 사용 건 수가 평균적으로 281.19건/1개월
평균 문자량	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 총 문자 사용 건 수가 평균적으로 21.18건/1개월
평균 통화대상자 수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 총 통화대상자 수가 평균적으로 21.47명/1개월
평균 문자대상자 수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 총 문자대상자 수가 평균적으로 9.41명/1개월
SNS 사용횟수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 총 SNS 사용횟수의 표준화된 값이 평균적으로 -0.05
소액결제 사용횟수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 소액결제 사용 횟수가 평균적으로 5회/1개월
소액결제 사용금액	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 소액결제 사용 금액이 평균적으로 10750원/1개월
최근 3개월 내 요금 연체 비율	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 핸드폰 요금 연체 여부가 있는 사람의 비율이 0.02 (=2%)

변수	해석 예시
야간상주지 변경 횟수	강남구 개포1동에 거주하는 사람들의 최근 36개월 기준 야간상주지의 변경횟수가 평균적으로 2.49회/1년
주간상주지 변경 횟수	강남구 개포1동에 거주하는 사람들의 최근 36개월 기준 야간상주지의 변경횟수가 평균적으로 3.24회/1년
평일 총 이동거리 합계	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 평일 총 이동거리 합계가 평균적으로 455Km/1개월 평일
휴일 총 이동거리 합계	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 휴일 총 이동거리 합계가 평균적으로 440.74Km/1개월 휴일
평일 총 이동 횟수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 평일 총 이동 횟수가 평균적으로 23.14회/1개월 평일
휴일 총 이동 횟수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 휴일 총 이동 횟수가 평균적으로 13.24회/1개월 휴일
집추정 위치 평일 총 체류시간	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 집추정 위치의 평일 총 체류시간이 평균적으로 17045.85분/1개월 평일
집추정 위치 휴일 총 체류시간	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 집추정 위치의 휴일 총 체류시간이 평균적으로 5399.15분/1개월 휴일

변수	해석 예시
평균 출근 소요시간	강남구 개포1동에 거주하는 사람들의 최근 1개월 기준 주중 출근 1회당 출근 소요시간의 한달 평균이 평균적으로 84.69분/출근 1회
평균 근무시간	강남구 개포1동에 거주하는 사람들의 최근 1개월 기준 주중 근무 1일당 근무시간의 한달 평균이 평균적으로 511.25분/평일 하루
지하철이동일수 합계	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 총 지하철 이용일수가 평균적으로 11.88일/1개월
동영상/방송 서비스 사용일수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 동영상/방송 서비스 사용일수가 평균적으로 71.25일/1개월
유튜브 사용일수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 유튜브 사용일수의 표준화된 값이 평균적으로 69.84일/1개월
넷플릭스 사용일수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 넷플릭스 사용일수의 표준화된 값이 평균적으로 35.23일/1개월

변수	해석 예시
게임 서비스 사용일수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 게임 서비스 사용일수가 평균적으로 25.14일/1개월
금융 서비스 사용일수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 금융 사용일수가 평균적으로 45.72일/1개월
쇼핑 서비스 사용일수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 쇼핑 사용일수가 평균적으로 12.95일/1개월
배달_배달 서비스 사용일수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 배달 서비스 사용일수가 평균적으로 44.92일/1개월
배달_브랜드 서비스 사용일수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 배달 브랜드 서비스 사용일수가 평균적으로 11.46일/1개월
배달_식재료 서비스 사용일수	강남구 개포1동에 거주하는 사람들의 최근 3개월 기준 1개월 평균 배달 식재료 서비스 사용일수가 평균적으로 7.57일/1개월

2 데이터 공개방법

서울 시민생활 데이터는 **열린데이터광장**(data.seoul.go.kr)을 통해,
통신사 자체 검증 후 **월단위(익월 20일) 개방**

공간 단위 25개 자치구, 424개 행정동

작성 정보 10개 관심집단수, 29개 통신정보

작성 단위 성, 연령, 1인가구, 시민전체

작성 주기 월 단위 (자체 검증 후 익월 20일 개방)

행정동 코드	자치구	행정동명	성별	연령대	총인구	1인가구수	커뮤니케이션이 적은 집단	평일 외출이 적은 집단	---	외출-커뮤니케이션이 모두 적은 집단
1123068	강남구	개포1동	01	20	616	95	25	17	---	0
1123068	강남구	개포1동	01	25	547	97	18	11	---	0
1123068	강남구	개포1동	01	30	482	138	6	3	---	0
1123068	강남구	개포1동	01	35	366	95	10	5	---	0
1123068	강남구	개포1동	01	40	302	32	4	3	---	0
1123068	강남구	개포1동	01	45	515	57	7	4	---	0
1123068	강남구	개포1동	01	50	658	61	18	4	---	1
1123068	강남구	개포1동	01	55	638	60	4	1	---	1
1123068	강남구	개포1동	01	60	475	46	10	1	---	0
1123068	강남구	개포1동	01	65	346	50	6	1	---	0
1123068	강남구	개포1동	01	70	195	26	3	1	---	1
1123068	강남구	개포1동	01	75	246	60	1	0	---	0

(참고) 제공기간: 2022년 1월~

❸ 분석 및 시각화 예시

Q1. 20대 초기 남성중 재정 상태에 관심이 필요한 집단이 많이 거주하는 행정동은 어디일까요?

- A1: 20대 초기(20~24세) 남성중 재정 상태에 대한 관심 집단이 많이 거주할 것으로 추정되는 행정동은 양천구 신정3동(추정된 1인가구의 수는 69.8)입니다.

Q2. 20대 초기 남성중 재정적으로 양호한 사람이 많이 거주하는 행정동은 어디일까요?

- A2: 20대 초기(20~24세) 남성중 재정 상태에 대한 관심이 필요한 집단이 적게 거주할 것으로 추정되는 행정동은 중구 을지로동(추정된 1인가구의 수는 0.2)입니다.

➡ 추가적으로 재정 상태에 대한 관심 집단의 지역적 분포(행정동별 분포)는 다음의 python 시각화 코드를 통해 확인하실 수 있습니다.

Python 코드 예시 1

1인가구 시각화를 위한 파이썬 코드

- 지수로 산출된 관심 1인가구 수를 행정동별로 시각화하는 코드
- 연령 20~24세, 남성 중 평일 외출이 적은 1인가구 수를 시각화

```
# 필요 파일
# - 1인가구.csv, data.shp,
import os
os.chdir(os.path.dirname(os.path.abspath(__file__)))
# OS가 window인 경우 geopanda 설치 오류가 발생하는 경우
# 필수 의존성 패키지 설치를 확인
# 1. pyproj, 2. Shapely,  3. GDALL, 4.Fiona

import geopandas as gpd
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib as mpl

mpl.rc('font', family='Malgun Gothic') # 한글 폰트 적용
plt.rcParams["figure.figsize"] = (20, 10) # 차트 사이즈
mpl.rcParams["axes.unicode_minus"] = False
```

```
"""1인가구 지수 데이터"""
# 데이터 형식의 통일
data = pd.read_csv('1인가구.csv', encoding='cp949')
n = data.shape[0]
adm_cd = []
for i in range(n):
    adm_cd.append(str(data['행정동코드'][i]))
data['adm_cd'] = adm_cd



"""행정동 위치 파일과 결합"""
geo = gpd.read_file('data.shp')
df_geo = geo.iloc[:,[2,9]]
rdata = pd.merge(data, df_geo, on = 'adm_cd')
data_merge = gpd.GeoDataFrame(rdata, crs="EPSG:4326", geometry="geometry")
```

```
"""시각화: 20~24세, 남성 평일 외출이 적은 집단"""
# 추출
data_tmp = data_merge[data_merge['연령대'] == 20][data_merge['성별'] == 1]
data_tmp.head()

# 시각화,  column 에 시각화할 필드명을 입력
fig = plt.figure()
data_tmp.plot(column='평일 외출이 적은 집단',
              legend=True,
              cmap='YlGn',
              edgecolor='k',
              legend_kwds={'label': '명'})
plt.axis('off')
plt.tight_layout()
plt.savefig('example_1.png')
plt.show()
```

시각화: 20~24세, 남성 평일 외출이 적은 집단

* 행정동 코드(geometry) 2020년 7월 1일 기준

Python 코드 예시 2

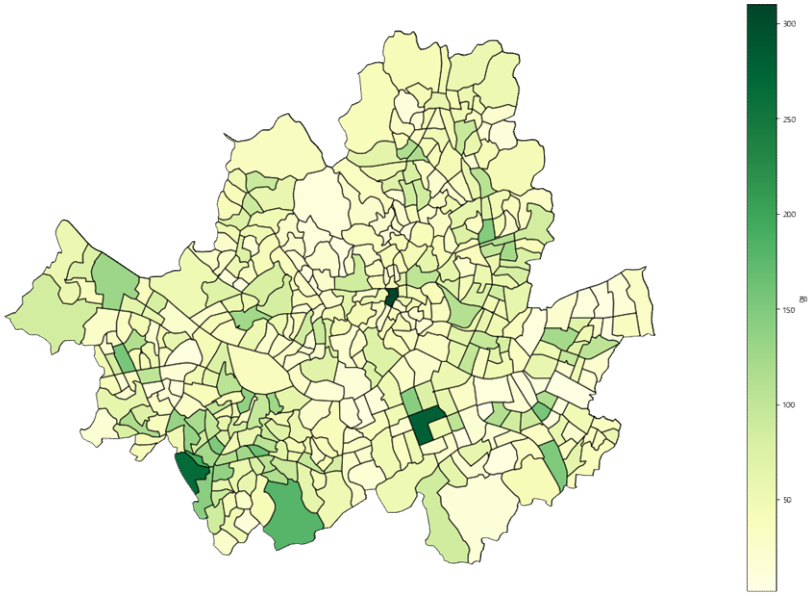
1인가구 시각화를 위한 파이썬 코드

- 연령대를 병합하여 시각화하는 파이썬 코드

```
# 연령대의 병합
# 20-29: 20대 (초기청년층)
# 30-39: 30대 (후기청년층)
# 40-49: 40대 (중년층)
# 50-64: 50대 (장년층)
# 65: 65세 이상 (노년층)
data_merge['연령대1'] = data_merge['연령대']

idx = (data_merge["연령대"] == 25)
data_merge["연령대1"][idx] = 20

idx = (data_merge["연령대"] == 35)
data_merge["연령대1"][idx] = 30

idx = (data_merge["연령대"] == 45)
data_merge["연령대1"][idx] = 40

idx = (data_merge["연령대"] == 55)
data_merge["연령대1"][idx] = 50

idx = (data_merge["연령대"] == 60)
data_merge["연령대1"][idx] = 50

idx = (data_merge["연령대"] > 60)
data_merge["연령대1"][idx] = 65

 colname_ = list ( data_merge.columns[5:17] )
data_groupby = data_merge.groupby(["adm_cd",'연령대1'])[colname_].sum()
data_groupby = data_groupby.reset_index()
data_groupby = pd.merge(data_groupby, df_geo, on = 'adm_cd')
data_groupby = gpd.GeoDataFrame(data_groupby, crs="EPSG:4326",
                                geometry="geometry")
```

```
# 중년층 중 휴일 외출이 적은 집단
data_tmp = data_groupby[data_groupby['연령대1']==40]
fig = plt.figure()
data_tmp.plot(column='휴일 외출이 적은 집단',
              legend=True,
              cmap='YlGn',
              edgecolor='k',
              legend_kwds={'label': '명'})
plt.axis('off')
plt.tight_layout()
plt.savefig('example_2.png')
plt.show()
```

시각화: 중년층 (40~49세) 휴일 외출이 적은 집단



* 행정동 코드(geometry) 2020년 7월 1일 기준

4. 유의사항 및 한계점

1 집단들의 시계열적 해석에서의 주의사항

유튜브와 넷플릭스 사용일수, 그리고 SNS 사용횟수는 다른 서비스와 달리 수집된 시점의 전체 데이터 평균과 표준편차로 표준화되어 지수 형태(z-score)로 제공된다. 따라서 수집된 시점에 따라 변수의 평균과 표준편차가 달라 지수의 계산방식이 달라지므로 해당 변수들의 시점 간 비교가 어렵다.

2 소득 변수의 의미

본 데이터 분석에서 소득 변수는 통계청의 건강보험 연말정산 보수총액 코드를 이용하였는데, 해당 변수는 근로소득만을 고려하므로 소득 관련 지수의 해석에서 유의해야한다.

3 이동거리 및 위치 정밀도 관련

이동거리가 5km 단위로 구간화되어 있어, 0~5km 미만의 이동거리를 갖는 사람이 모두 이동거리가 0km인 것으로 변수가 구성되었다.

체류시간 및 이동 건 수 측면에서, 해당 변수들은 야간 상주지(행정동)를 기준으로 측정되므로, 행정동 내에서의 적은 거리의 이동이나 외출은 알기 어렵다는 한계가 있다.

4 노년층 후기(75세 이상)의 데이터 수

다른 연령층에 비해 데이터의 수가 적으므로, 통계적 결과 해석에 있어서 반드시 유의해야 한다.

FAQ

❶ 데이터 소개 및 구축 방법

Q1. '서울 시민생활 데이터'란 무엇입니까?

서울시와 SK텔레콤이 통계청 데이터와 통신데이터 가명결합을 통해 추정한 서울 행정동 단위 성, 연령별 1인가구의 생활특성을 엿볼 수 있는 데이터로 커뮤니케이션지수, 이동지수, 재정지수 등 10개 지수와 29개 통신정보(통화량, 외출횟수, 요금연체여부, 모바일 서비스 사용량 등)를 말합니다.

Q2. '서울 시민생활 데이터'를 개발한 이유는 무엇입니까?

시민생활 데이터 개발 목적은 1인가구의 삶의 질, 생활특성을 분석하여 1인가구 정책을 지원하기 위함입니다. 그동안 인구총조사, 주민등록인구통계, 설문조사 등을 통해 정책개발을 하고 있었으나, 공간적 혹은 시간적 제약이 많아 좀 더 시의성 있고 해상도 높은 데이터가 필요하여 개발했습니다.

* 공간해상도: 자치구 단위 → 행정동 단위
* 시간해상도: 1년 주기 생산 → 매월 생산

Q3. 왜 SKT와 협력을 추진하게 되었습니까?

통신사는 각 가입자의 거주지나 근무지를 파악하기 쉽고, 관심사나 재정상태 파악에도 용이한 데이터를 갖고 있습니다. 이런 이유로 통신데이터를 통해 1인가구의 수와 생활특성을 파악하고자 하는 첫 번째 시도를 하게되었습니다.

Q4. SKT가 제공하는 통신정보 데이터란 무엇입니까?

개인정보 식별이 안되는 통계화된 데이터로 휴대폰 가입자의 위치정보를 이용해 추정한 주·야간 상주지 정보, 휴대폰요금 연체와 소액결재 등 재정상태 관련 정보, 다양한 서비스 이용량에 대한 집계정보를 말합니다.

FAQ

Q5. 휴대폰을 가지고 있지 않는 아동이나 노인 연령대에 대한 추정은 어떻게 진행되었나요?

방송통신위원회 통계에 의하면 10세 이상 아동의 휴대폰 보급율은 70% 이상, 70세 이상 노년층의 휴대폰 보급률도 90% 이상입니다. 시민생활 데이터 즉, 1인가구 분석대상은 20세 이상으로 개발되었기 때문에 휴대폰 보급률 영향은 크지 않습니다.

Q6. 행정동 집계 기준이 어느 시점으로 되어있나요?

데이터는 서울시 열린데이터 광장을 통해 매월 20일경 공개되며, 기준월 직전 최근 3개월 평균값입니다.

Q7. 1인가구는 실거주인지 혹은 공부상 등록된 1인가구인지요?

통계청의 등록센서스(1인가구)는 공공데이터로 추정된 실거주지, 행안부의 주민등록인구통계(1인세대)는 등록된 주소지 기반의 1인가구(세대)이며, 이번에 발표하는 서울 시민생활 데이터의 1인가구는 심야시간대(새벽 1시~6시) 휴대폰 위치 기반으로 추정한 인구입니다.

Q8. 연령대 분류 기준은 무엇인가요?

1인가구 정책관련 각 분야 전문가들의 의견을 수렴하여 20대(초기청년층), 30대(후기청년층), 40대(중년층), 50-64세(장년층), 65-74(초기노년층), 75세이상(후기노년층) 6개 연령층으로 구분하여 분석하였으며, 시민생활 데이터의 기본 연령대는 5세 단위로 생산, 개방할 예정입니다.

FAQ

Q9. 결측치 처리는 어떤 방식으로 진행하셨나요?

통신데이터에서 결측치는 거의 없습니다. 특정 서비스를 사용하지 않은 경우 관측값에는 0의 값을 가지며, 분석시에도 0 값으로 이용하였습니다. 데이터 오류 문제로 결측치가 있는 경우에는 값을 확인하고 분석에서 제외하였습니다.

Q10. 조작적 정의의 기준은 무엇인가요?

관심집단인 외출이 매우 적은 집단(전체), 외출이 매우 많은 집단, 외출-커뮤니케이션이 모두 적은 집단(전체)을 구분하기 위해 몇 가지 조건을 확인했습니다. 이 조건은 매뉴얼 본문 페이지를 참조바랍니다. 사용한 임계값은 분산분석, 회귀분석의 변수선택 등의 방법을 통해 집단의 특성이 가장 잘 구분되는 값을 정하고 전문가의 의견수렴을 거쳐 확정하였습니다. 특히 이동거리에 대한 조건은 관측된 이동거리의 오차 문제로 다양한 조건을 복합적으로 설정하였습니다.

Q11. 동영상, 방송서비스 대표 서비스는 무엇이 있을까요?

웨이브, 와차, 티빙 등입니다.

Q12. 금융, 쇼핑, 배달 대표 서비스는 무엇이 있을까요?

금융서비스는 증권, 은행, 가계부 등이며 쇼핑은 의류나 종합 쇼핑서비스, 배달관련 대표 서비스는 일반적인 배달서비스와 치킨/피자 등 특정 브랜드의 배달서비스, 유기농 식품의 배송 서비스 등 식재료 배달서비스 입니다.

FAQ

❷ 활용 및 배포 방안

Q1. 구체적으로 어떤 정책에 활용할 예정입니까?

분석을 통해 1인가구는 성, 연령, 거주지역에 따라 다양한 특성을 보이고 있어, 기존 설문조사에 기반한 서울시의 1인가구 4대 정책과 병행하여 고립, 주거안심 등 1인가구 특성에 맞춘 핀셋정책에 활용될 예정입니다.

Q2. 시민이 어떻게 활용할 수 있습니까?

이번에 개발된 데이터는 서울시 빅데이터캠퍼스를 통해 개방될 예정입니다. 데이터의 쉬운 사용을 위해 매뉴얼을 개발하고 간단한 활용사례를 함께 배포할 예정입니다.

Q3. 시민생활 데이터의 배포 주기는 어떻게 됩니까?

공동연구 기관인 SKT텔레콤과 서울시립대의 자체 검증 후 전월 데이터를 매월 20일경 배포할 예정입니다.

❸ 유사 통계, 데이터와의 비교

Q1. '시민생활 데이터'는 통계청 등록센서스 1인가구, 행정안전부의 주민등록상 1인세대와 어떤 차이가 있습니까?

기존에 발표되고 있는 통계청 등록센서스에서의 가구는 1인 또는 2인 이상이 모여서 취사, 취침 등 생계를 같이 하는 생활단위를 의미하며, 행안부의 주민등록은 전·출입 신고에 위해 등록된 집주소기반 1인세대를 의미합니다. 이번에 발표하는 서울 시민생활 데이터의 1인가구는 심야시간대(새벽 1시~6시) 휴대폰 위치 기반으로 통계모형을 적용해 추정한 인구입니다. 각 데이터마다 작성목적, 작성기준이 달라 데이터간 특징과 차이가 있습니다. 따라서 다른 통계와 수치적으로 직접 비교하기보다는 상호 보완적으로 활용할 필요가 있습니다. 또한, 1인가구의 숫자보다는 통신정보를 활용해 첫 번째 시도한 1인가구에 대한 커뮤니케이션 지수, 재정지수 등 그동안 없었던 새로운 데이터를 정책과 연구 등에 활용되었으면 합니다.

FAQ

4 분석결과

Q1. 분석시점이 2020년 11월 기준인데, 코로나 영향으로 이동이나 여가, 생활패턴에 영향을 받지 않았을까요?

최초 분석한 소득, 연락, 이동 등 많은 생활패턴이 코로나의 영향을 받았을 것으로 예상됩니다. 특히, 경제활동이 위축되고 외부 여가활동이 코로나 이전보다 줄어들었을 것으로 예상됩니다. 비교 시점 간 편향성이 통계적으로 확인되지는 않았지만 1인가구 추정, 생활실태 분석을 위해 개발한 모형과 결과 값은 코로나 기간의 특수성을 감안하여 사용해야 합니다. 향후, 매년 분석 모형을 지속적으로 학습하여 과거 데이터의 편향성을 줄여나갈 계획입니다.

Q2. 1인가구가 유사 통계와 다른 값을 갖고 있었습니다. 신뢰할만한 새로운 데이터 인가요?

어떤 데이터가 맞다 틀리다 할 수 없습니다. 유사 통계, 데이터간 작성방법이 다르고 기준도 다르기 때문입니다. 1인가구 수에 대한 신뢰성에 앞서 1인 가구를 식별하는 방법이 선행되어야 할 것 같습니다. 1인가구가 주민등록기준이 아닌 통신데이터 특성으로 정의하고 생활실태에 근거하여 만들어진 데이터라는 특징이 있습니다. 통신데이터와 통계청 인구총조사 자료를 결합하여 1인가구 추정방법을 정교화하고 정확도를 높이고자 했습니다. 향후 통신데이터 이외 소비데이터, 소득·부채 데이터 등 민·관 다종 데이터를 추가 결합할 예정으로 1인가구 추정 정확도와 개선효과가 있을 것 기대합니다.

FAQ

Q3. 로지스틱 회귀 모형을 이용해 1인가구일 확률로 1인가구를 추정하셨다 했는데, 최선의 모형일까요?

특정 구역의 1인가구 수를 추정하기 위해서는 회귀모형을 사용하는 것이 분석과 결과해석에 편리합니다. 하지만 이런 방식으로 만들어진 모형은 개인단위의 문제분석에 활용이 어렵습니다. 반면 로지스틱 회귀 모형을 이용할때 개인의 1인가구 확률을 개별적으로 추정하여 총합을 산출하는 방식은 향후 모형개발 및 관리의 유연성 측면에서 장점이 있습니다. 특정 지역의 관측가능한 인구수의 총합을 알아야 한다는 단점이 있지만, 통신데이터의 보정을 통해 이 문제를 비교적 쉽게 해결할 수 있습니다. 한편 2020년 11월 분석데이터를 기준으로 1인가구 추정방법 선택을 위한 다양한 기계학습 방법을 비교하였으나 모형간 큰 차이를 발견할 수 없었습니다.

Q4. 세분화된 1인가구집단을 예측하는 로지스틱 회귀모형의 정확도는 어떻게 되나요?

적합모형의 목적은 개개인이 관심집단에 속하는지 판별하는 것이 목적이 아니라 특정 집계단위 내에서의 관심집단의 수, 즉 추정된 수의 기대값을 추정하는 것이 목적입니다. 따라서, 적합된 모형의 정확도는 기준치(baseline)가 되는 전체 데이터에서의 관심집단의 비율과 거의 일치합니다.

Q5. '야간상주지'와 '주간상주지'를 어떻게 추정하나요?

통신데이터의 기지국을 기준으로 특정 시간대에 위치가 변하지 않았던 장소를 상주지라고 합니다. 주간상주지와 야간상주지는 주간 시간대(11시~15시)와 야간 시간대(01시~06시)에 위치의 변화가 가장 작았던 장소를 말하며, 주간에 특정장소에서 일하고 야간에 집에서 쉬는 사람을 기준으로는 주거지와 근무지로 해석할 수 있습니다. 매 월 1일에 전 월 1개월 간의 위치를 기준으로 추정하고 있습니다.

FAQ

Q6. 평일과 휴일 이동거리는 어떻게 산출된 건가요?

야간 상주지역 외 타 행정동으로 이동한 경우를 직선 거리로 추정합니다.

Q7. 외출도 애매합니다. 집 인근 편의점, 친구 집 등에 가도 외출로 잡히나요?

외출은 미리 정의한 야간상주지에서 다른 곳으로 이동(기지국 간 이동)을 확인할 수 있을 때 식별할 수 있습니다. 야간상주지에서 가까운 곳으로 이동(기지국 내 이동)한 집 밖 외출은 식별할 수 없습니다.

Q8. 배달서비스 분석시점과 현재는 차이가 있어 보입니다. 최근 배달서비스 사용량 분석도 공개 가능할까요?

오늘 공개된 데이터는 서울시 열린데이터 광장을 통해 매월 공개 예정으로 시민 누구나 제약 없이 사용 가능합니다.

Q9. SKT에서 추정한 20대 1인가구 밀집지역과 등록센서스 20대 1인가구 밀집지역 분포가 유사한가요?

두 데이터간 행정동별 20대 상관도를 보았을 때 0.9 이상으로 유사한 분포를 보였습니다. 다만, 40대 분포는 좀 더 차이가 가장 커서 다종 데이터와 결합해 추가 연구가 필요합니다.

- 상관계수 범위: –1에서 1사이, ±0.7이상(강한 선형상관), ±0.3 ~±0.7(뚜렷한 선형상관), ±0.1~±0.3(약한 선형상관), ±0.1(무시할 수준의 선형상관)

Q10. 40대 1인가구와 일반적인 40대가구 거주지 분포가 차이가 나는 이유는 무엇이라 생각하나요?

일반적인 40대 가구의 경우 주거시설이 많은 지역에 주로 거주합니다. 그러나 삶이 상대적으로 안정적이지 못한 40대 1인가구는 주거 상황이 열악하더라도 일자리를 구할 수 있는 기회가 많거나 상대적으로 주거비용이 적게드는 지역을 선호하기 때문인 것으로 추정됩니다.

FAQ

Q11. 본 분석결과에서 40대가 가장 취약해 보이는데, 서울시에서의 지원 정책은 마련되어 있나요?

바로미터는 아니지만, 소액결제 금액이나 휴대폰 요금 연체율이 중·장년층(40~64세)에서 10%대를 상회하는 등 청년층과 노년층보다 재정적 위기 상황일수 있다는 생각이 들었습니다. 제도적으로 도움을 줄 수 있는지 면밀히 살펴 볼 필요가 있으며, 오히려 원래 살던 지역에 계속 살고 싶어하는 경향이 더 강한 중장년, 특히 노년을 위한 공공주택을 공급하거나 공동체 생활을 지원하는 프로그램 등이 필요하지 않을까 합니다.

Q12. 1인가구의 소비, 부채 상태 등도 추가로 분석할 계획인가요?

1인가구의 소비와 부채를 통한 재정위기 파악은 서울시의 중장년 1인가구 정책에서 매우 중요한 일입니다. 카드사와 신용정보를 활용한 1인가구의 소비, 경제활동 상태를 추가로 결합분석할 계획이며, 통신데이터로 파악하기 어려운 1인가구 경제활동 특성을 면밀하게 분석할 수 있을 것으로 기대하고 있습니다.

Q13. 세분화된 1인가구 집단을 예측하는 로지스틱 회귀모형의 정확도는 어떻게 되나요?

적합된 모형의 목적은 개개인이 집단의 속하는지 여부를 맞추는 것이 아닌, 어떤 집계 단위 내에서의 관심집단의 수, 즉 추정된 수의 기댓값을 맞추는 것이 목적입니다. 따라서, 적합된 모형의 정확도는 기준치(baseline)가 되는 전체 데이터에서의 관심집단의 비율과 거의 일치합니다.

FAQ

Q14. 관심집단을 집계할 때, 관심집단에 속할 확률을 더한 것이 아니라, 10% 또는 2%의 기준을 설정한 뒤 집계한 이유는 무엇일까요?

결합 데이터는 특정 시점에서 수집된 데이터이므로, 결합 데이터만으로는 계절적 변화와 같은 시계열적으로 해당 관심집단에 해당하는 사람의 숫자가 어떻게 변하는지 분석하기 어렵다는 한계가 있습니다. 따라서, 계절적 요인에 영향을 받지 않고 집단을 집계하는 방법이 필요했습니다.
이를 위해 개발한 관심집단 모형으로부터 계산할 수 있는 관심집단에 속할 확률을 각 개인의 점수로 고려하였습니다. 그리고 점수가 높은 10% (혹은 2%)의 인구가 서울시에 현 시점에 어떻게 분포하였는지 확인하는 방법을 본 분석에서 활용하였습니다. 이러한 집계 방식을 통해, 추후 생산될 데이터로부터 점수가 높은 동일한 10% (혹은 2%)의 서울 시민의 지역적 분포를 시계열적으로 확인하는 것이 가능해졌습니다.

Q15. 미추정 인구수가 제공되지 않는 변수의 이유는 무엇인가요?

통신데이터 자체가 SKT 가입자를 대상으로 집계된 것으로, 기본적인 통신 서비스인 통화와 문자, 데이터 등을 사용하지 않는 사용자가 매우 적었습니다. 따라서, 이러한 변수들에 대해서는 식별 가능성의 이유로 미추정 인구수를 제공하지 않고 있습니다.

FAQ

Q15. 둔촌1동에 대한 데이터 결과가 상식적으로 잘 이해가 되지 않는데, 원인이 무엇인가요?

둔촌1동의 경우 재개발 사업이 추진되고 있는 곳으로 일반적인 행정동의 통신데이터 분포와 매우 다르게 나타날 가능성이 있습니다. 1인가구수의 추정, 각종 지수의 산출등 많은 추정값들이 통신데이터의 특성에 의존하고 있어 특이한 사건으로 인해 데이터의 특성이 달라질 지역의 추정정확도 및 자료 신뢰성의 문제를 가지고 있습니다. 이런 이유로 극단적인 패턴을 가지는 지역에 대해서 자료해석이 신중을 기할 필요가 있습니다.

Q16. 게임, 금융과 같은 서비스의 사용량에서 이용일수로 이를 정의하는 이유는 무엇인가요?

사용 횟수를 이용할 경우 실수로 잠깐 이용한 경우까지 모두 사용횟수로 고려가 됩니다. 그리고 유사한 서비스 간 이용 회수의 차이가 존재하는 문제도 있습니다. 예를들어 게임의 경우 짧게 자주 이용하는 것이 유리한 게임과 장시간 지속적으로 잉요하는 것이 유리한 게임이 있습니다. 이러한 차이를 감안하여 사용일수로 사용량을 정의하고 있습니다.