# 2025년 새싹 해커톤(SeSAC Hackathon) AI 서비스 기획서

| 팀명 | SBAI |
|---|---|
| 팀 구성원 성명 | 장원석, 권용진, 김동준, 이영제 |

## 1. AI 서비스 명칭

| **통계 검색 챗봇, 이지스탯 Q** |
|---|

## 2. 활용 인공지능 학습용 데이터

| | 활용 데이터명 | 분야 | 출처 | 목적 |
|---|---|---|---|---|
| 1 | 주제별 통계목록 | 공공(통계·행정) | 국가통계포털 | 테이블 스키마 검색 최적화 |
| 2 | 자연어 기반 질의(NL2SQL) 검색 생성 데이터 | 공공(공공데이터·질의응답) | AI HUB | 자연어 질의 -> SQL 학습 |
| 3 | 문장 유형(추론, 예측 등) 판단 데이터 | 일반 | AI HUB | 유형별 컨텐츠 생성 |

## 3. 핵심내용

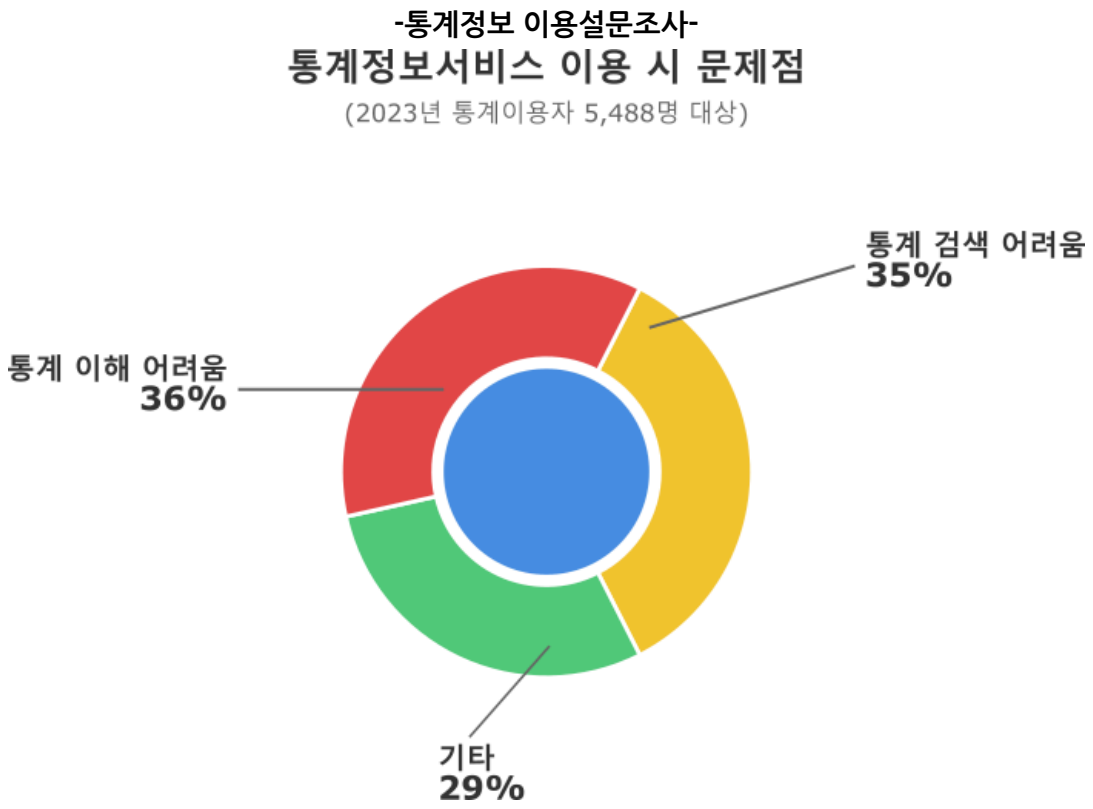
| 국가통계포털(KOSIS)에 있는 대량의 통계 데이터를 LLM을 활용, 전문가가 아닌 일반 통계이용자도 자연어로 쉽게 질의, 원하는 데이터를 빠른 시간 안에 검색하고 의도에 맞는 시각화와 데이터 분석 및 인사이트를 제공하며 나아가 통계를 기반으로 기사, 보고서, 블로그 유형의 컨텐츠를 요구에 맞게 생성, 시민들의 통계 접근성과 편의, 실용성을 높여 국가 데이터 공공성 상승에 기여한다. |
|---|

## 4. 제안배경 및 목적

**1.제안배경**

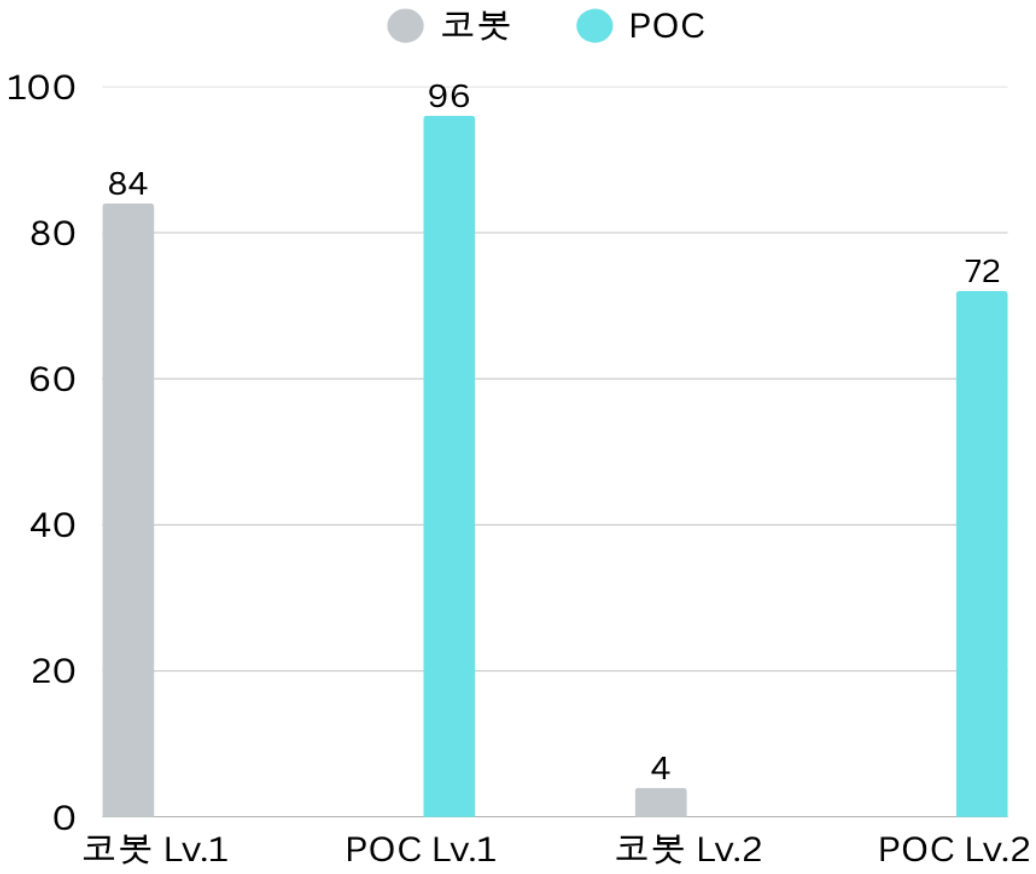
4차 산업혁명 시대에 데이터의 중요성은 두 말할 필요가 없다. 인구, 출산율, 취업률, 소득, 부동산, 범죄, 재난 등 우리 사회 곳곳의 문제와 변화는 통계를 통해 드러난다. 기업 경영, 마케팅, 학술연구, 창업에 이르기까지 수많은 의사결정이 통계에서 출발한다.

이러한 공공 통계 데이터를 쉽고 공정하게 이용하는 것은 '국민의 권리'에 가깝다. 그러나 현실에서 많은 사람들은 언론 기사 몇 줄에 의존해 통계를 수동적으로 '전달받는' 수준에 머물러 있다. 능동적으로 통계를 찾아보고 비교·분석하기에는 현재의 통계 서비스가 이용자 친화적이지 못하기 때문이다.




통계인재개발원의 「2023년 일반 통계이용자 대상 통계정보서비스 이용 시 문제점」 조사에 따르면, 일반 이용자의 35%가 '통계검색이 어렵다'고 응답했다. 통계표 구조와 용어, 배경지식이 있어야만 접근이 가능하다는 의미다. 이는 데이터 리터러시를 가진 소수만 공공 통계를 적극 활용하고, 다수 시민은 여전히 통계로부터 소외되는 디지털 격차 문제로 이어진다. 국가데이터처는 이러한 문제를 해결하기 위해 2021년, 규칙기반 검색 챗봇 '코봇'을 선보였지만, 복잡한 질의를 처리하는데 한계를 보였다.

-POC 테스트 결과-




본 프로젝트에서는 레벨별 자체 평가 데이터셋을 구축해 이지스탯 Q POC와 코봇의 성능을 비교·측정하였다. 그 결과, 다중 조회가 필요한 2단계(Lv.2) 이상 문제부터는 코봇이 의도한 결과를 안정적으로 반환하지 못해 실사용에는 상당한 제약이 있는 것으로 나타났다. 보다 유연한 자연어 이해와 질의응답을 지원하는 지능형 통계 서비스가 필요한 이유다.

**2.개발목적**

**“공공 통계데이터 접근성 및 공공성 향상”**

- 비전문가도 자연어 질의를 통해 **대화형 통계 검색 접근성 제공**
- 시각화 및 인사이트 분석 기능을 통한 정책·연구·교육 현장에서의 **통계 활용성 제고**
- 지역격차·인구구조·주거·노동 등 사회 현안을 데이터 기반으로 이해할 수 있는 시민·연구자·언론인의 데이터 저널리즘 실천 기반 마련 및 **생산성 향상**

## 5. 기술 스택

| | | |
|---|---|---|
| **AI/ML** | LangGraph | Multi-agent orchestration |
| | Text-to-SQL 모델 | 자연어 → SQL 변환 |
| | Intent Classification 모델 | 질문 유형 분류 |
| | LLM API | Gemini 2.5 Flash (프롬프트 기반) |
| **Backend** | Python FastAPI | API 서버 |
| | SQLite3/MySQL | 통계 데이터 저장 |
| | ChromaDB | 메타데이터 검색 |
| **Frontend** | Streamlit | 웹 UI |
| | Plotly | 시각화 자료 생성 라이브러리 |
| **Embedding** | Upstage Embeddings | Passage / Query |
| **Database** | Turso | 클라우드 SQLite |

## 6. 주요 기능

**1. 질문 분류**
LLM이 사용자 질문을 분석하여 6가지 시나리오 중 하나로 자동 분류한다.

**2. 테이블 및 스키마 검색**
벡터 DB 기반 유사도 검색을 통해 질문에 필요한 통계표를 자동으로 선택한다. 규칙 기반 로직을 결합하여 여러 테이블을 함께 선택할 수 있다.

**3. SQL 자동 생성**
자연어 질문을 SQL 쿼리로 자동 변환한다. 에러 발생 시 에러 메시지를 분석하여 최대 2회까지 재시도하며, 한국 통계 데이터 특성을 반영한 최적화된 쿼리를 생성한다.

**4. 데이터 조회 및 처리**
생성된 SQL을 실행하여 데이터를 조회하고, 파생 계산이나 다단계 분석이 필요한 경우 LLM이 추가 계산을 수행한다. 증가율, 성비, 평균값 등의 파생 지표를 자동으로 산출한다.

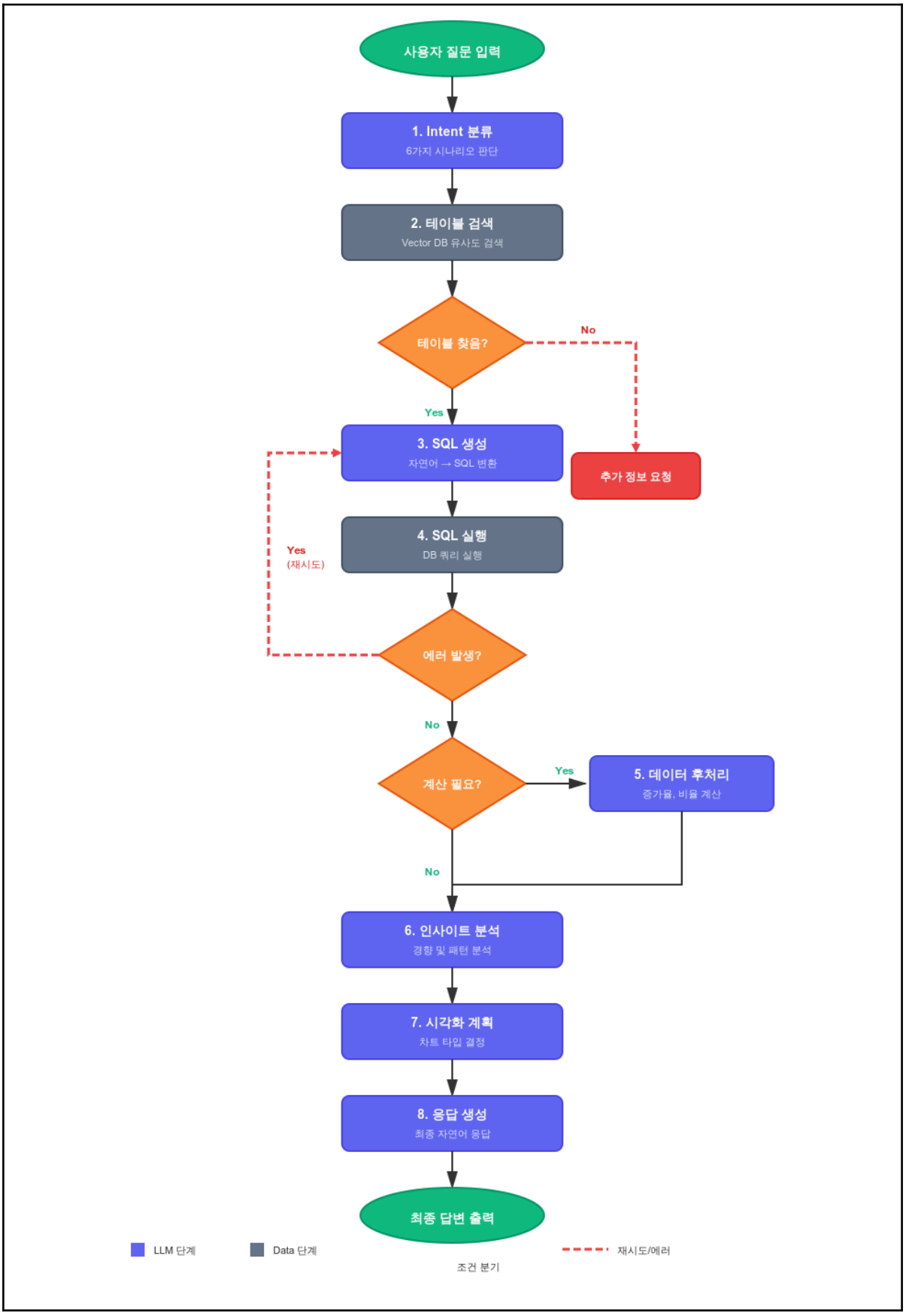
**5. 인사이트 분석**
조회된 데이터를 LLM이 분석하여 주요 경향과 패턴을 파악한다. 데이터의 증감 추세, 특이사항, 의미 있는 변화 등을 2-3문장으로 요약하여 제공한다.

**6. 시각화 자동 생성**
질문의 유형과 데이터 특성을 분석하여 적절한 차트 타입을 자동으로 선택한다. 시계열 데이터는 선 그래프, 비교는 막대 그래프, 비율은 파이 차트 등으로 시각화한다.

**7. 응답 생성**
자연어 답변, 조회 데이터, 인사이트, 시각화 차트를 통합하여 최종 응답을 생성한다. 사용자가 이해하기 쉬운 친절한 톤으로 완성도 높은 답변을 제공한다.

**8. 컨텐츠 생성**
기본 답변 외에 사용자가 선택할 수 있는 3가지 형식의 장문 컨텐츠를 생성한다. 기사체, 논문체, 블로그체 중 원하는 형식으로 통계 분석 결과를 재가공하여 제공한다.

## 7. 서비스 흐름도


사용자 질문 입력
1. Intent 분류
6가지 시나리오 판단
2. 테이블 검색
Vector DB 유사도 검색
테이블 찾음?
No
추가 정보 요청
Yes
3. SQL 생성
자연어 → SQL 변환
4. SQL 실행
DB 쿼리 실행
Yes
(재시도)
에러 발생?
No
계산 필요?
Yes
5. 데이터 후처리
증가율, 비율 계산
No
6. 인사이트 분석
경향 및 패턴 분석
7. 시각화 계획
차트 타입 결정
8. 응답 생성
최종 자연어 응답
최종 답변 출력
LLM 단계
Data 단계
재시도/에러
조건 분기

## 8. 시스템 아키텍처


사용자
Web/App 인터페이스
AI 에이전트 레이어 (LangGraph)
Intent Classifier
(Fine-tuned)
Table Selector
(Multi-label Model)
SQL Generator
(Fine-tuned Text-to-SQL)
Data Processor
(계산 및 처리)
Insight Analyzer
(인사이트 분석)
Content Generator
(컨텐츠 생성)
Data 레이어
Vector DB
(메타데이터 검색)
ChromaDB
Statistics DB
(KOSIS 100+ 테이블)
PostgreSQL

# 9. 평가 방법

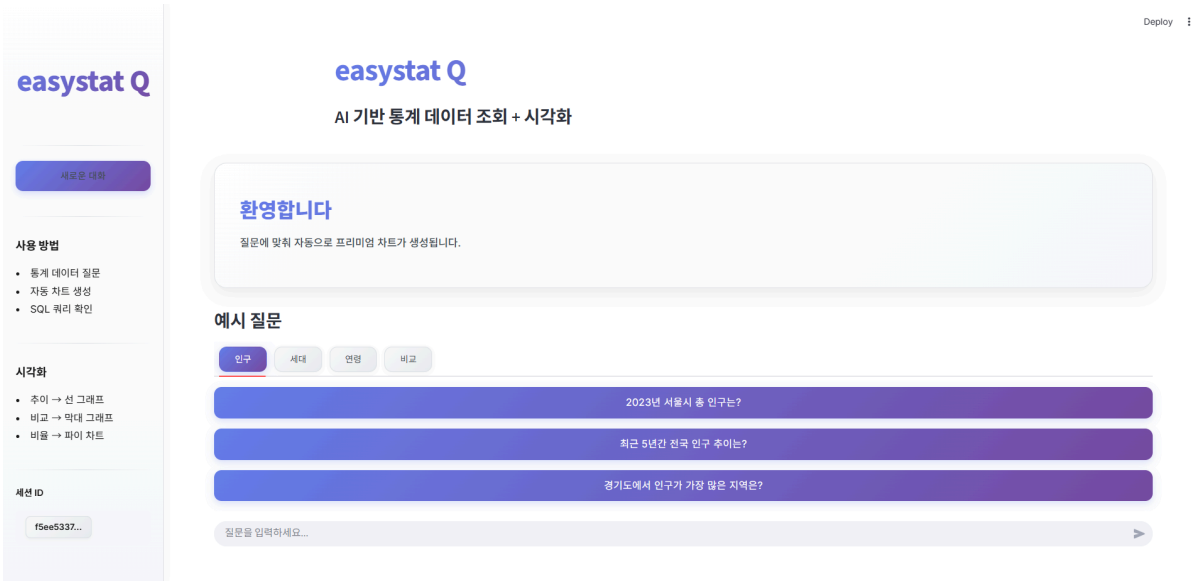
## 1. 레벨별 평가 데이터셋

| Lv | 테스트 범위 및 조건 | 문항 유형 | 문제 예시 |
|---|---|---|---|
| Lv.1 | 단일 테이블 단일 필드 값 | 추출 | 2016년 12월 대전광역시 총인구수는? |
| Lv.2 | 단일 테이블 다중 필드 값 | 범위, 조건, 합산, 개수, 다중 필드 선택 | 2024년 5월 서울특별시의 65세 이상 여자 인구는 얼마일까? |
| Lv.3 | 단일 테이블 필드값 가공 | 평균, 분산, 증감, 비율 | 2024년 7월 기준으로 서울이 전국 총인구에서 차지하는 비중을 알려줘. |
| Lv.4 | 다중 테이블 다중 필드 값 가공 | 여러 테이블 검색 | 2016년 5월 기준, 서울특별시 전체의 1세대당 남성 인구수는 얼마지? |

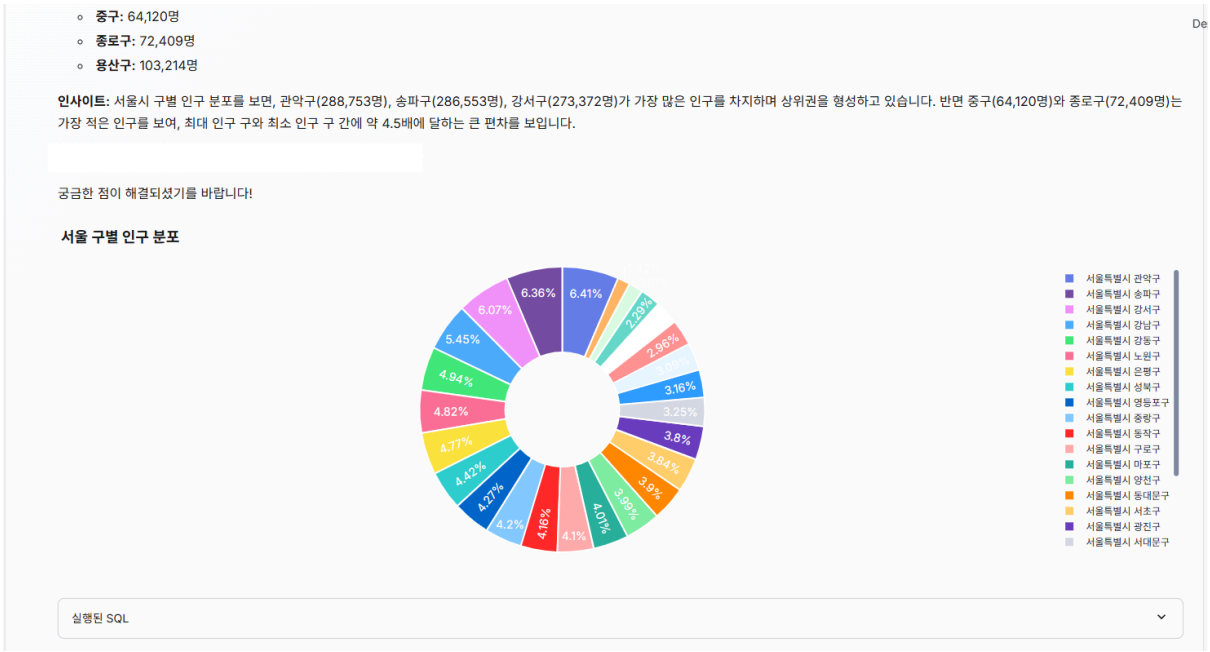
## 2. 평가 데이터셋 예시

| 유형 | 자연어 질의 | SQL Query | 정답 Label |
|---|---|---|---|
| 단순조회 | **서울특별시 2018년 6월 당시 세대수는 총 몇 세대였어?** | SELECT 값 FROM population_stats WHERE 행정구역='서울특별시' AND 년월='2018-06' AND 항목='세대수[세대]' ORDER BY id LIMIT 1; | 4241547 |
| 카운트 | **2018년 8월에 인구가 200만명 넘는 광역시도는 몇 개야?** | SELECT COUNT(*) FROM population_gender_stats WHERE 년월='2018-08' AND 항목='총인구수 (명)' AND 값 >= 2000000; | 9 |
| Lv.3 | **전국 기준으로 2022년 1월부터 6월까지 여자 인구 월평균은?** | SELECT AVG(값) FROM population_gender_stats WHERE 행정구역='전국' AND 항목='여자인구수 (명)' AND 년월 BETWEEN '2022-01' AND '2022-06'; | 2587725 8.6 |

# 10. UX/UI




**-서비스 메인 화면-**




**-차트 출력 화면-**

## 11. 기대효과

### 이지스탯 Q 개발 기대효과

| 항목 | 현행 | 이지스탯 Q 도입 후 | 개선 효과 |
|---|---|---|---|
| 통계 1건 검색 시간 | 5~40분/건 | 2~5분/건 | 약 50~80% 단축 |
| 통계 10건 연속 조회 | 60~180분 | 20~40분 | 약 60~80% 단축 |
| 기사·보고서 1편당 작업 시간 | 60~120분 | 30~60분 | 약 30~50% 단축 |
| 데이터 기반 문서 처리 건수 (2시간 기준) | 기준값 1.0 | 1.5~2.0배 | 생산성 1.5~2배 증가 |
| 다중 조건 질의 정답 도달률(Lv.4) | 20~40% | 70~85% | 정답률 2배 이상 향상 |
| 통계·그래프 포함 기사/콘텐츠 비율 | 전체의 10~20% | 30~40% | 비율 2배 전후 확대 |

*※ 위 수치는 국가통계포털(KOSIS) 사이트 활용 경험과 업무 단계 분석을 바탕으로 한 추정·목표 값이며, 프로젝트 테스트를 통해 검증·보정 예정.*

1.**통계 질의 1건당 검색 시간을 약 50~70% 단축**하여 검색 피로도를 낮추고 통계 접근성 향상 목표를 달성한다.

2.기사·보고서 등 데이터 기반 문서 작성 과정에서 통계 수집·그래프 작성 시간을 줄여, **전체 작업 시간을 약 30~40% 단축**하고 처리 건수를 1.5~2배까지 늘리는 효과를 기대한다.

3.인구·지역격차·주거·노동 등 사회 현안을 공공 통계로 쉽게 이해할 수 있어, **공공 통계 접근성 향상**과 함께 데이터 저널리즘 및 팩트 기반 보도 활성화에 기여할 것으로 기대한다.