빅데이터 분석 결과 보고서

신도시의 성공적인 병의원 개원을 위한 입지 분석 서비스

IPAD-1

곽동영, 권현준

목차

1. 분석 배경 및 기획의도
   1. 분석 배경
   2. 기획 의도
2. 분석 목표
3. 분석 데이터
   1. 데이터 수집
   2. 데이터 정합성 검사
   3. 데이터 전처리
4. 분석 과정 및 결과
   1. 서울 매출 분석 결과
   2. 지역 매출 분석 결과 및 시각화
5. 활용방안
6. 참고문헌

1. 분석 배경 및 기획의도

1.1 분석 배경

전국 15,000여개의 치과 의원 중 33%에 해당하는 5,000개의 치과 의원이 서울에 밀집되어 있는 것으로 나타났다. 인구 10만명 대비 치과의원 수에서도 2023년 12월 기준 서울은 10만명 당 51.95개를 기록하여 45.08개의 광주, 40.54개의 부산과 큰 차이를 보이며 이미 포화된 상태라고 할 수 있다. 이에 따라 서울에서의 개원이 아닌 다른 지역에서 개원을 하려는 의사들이 증가하고 있다.

개원을 하기 위해서는 많은 자금이 들어가기 때문에 여러 사항들을 고려하지 않고 개원하였다가 폐원할 시 손해가 상당히 크다고 할 수 있다. 따라서 개원 시 주변 환경에 대한 조사는 필수라고 볼 수 있다.

1.2 기획 의도

서울을 벗어나 신도시에서 개원하려는 의사들에게 매출 분석 서비스를 개발하여 도움을 줄 수 있도록 한다. 입지에 대한 정보와 진료 과목의 평균 데이터를 분석하여 예상 매출을 제공하여 개원하는 데 도움을 제공한다.

2. 분석 목표

-서울의 자치구별/연도별 데이터와 신도시의 지역 정보 데이터를 활용하여 해당 지역의 치과 의원 개원 시의 매출 분석 및 예측

3. 분석 데이터

3.1 데이터 수집

- 서울 지역 데이터

서울의 여러 가지 데이터 중에서 개원 입지와 관련되어 있다고 생각하는 데이터들을 수집하였다. 인구, 연령별 인구, 유동 인구, 생활 인구, 치과 의원 개수, 버스 정류장 수 등의 데이터로 환경에 관한 데이터를 구성하였고 매출, 임대료, 소득수준을 경제와 관련된 데이터로 구성하였다. 위의 데이터는 자치구별로 2019년부터 4개년의 데이터를 사용하였다.

> 인구, 연령별 인구\_ 통계청

> 소득수준 \_ 국세통계포털

> 유동 인구, 생활 인구, 버스 정류장 수, 매출, 임대료 \_ 서울데이터광장

> 치과 의원 개수 \_ 보건의료빅데이터개방시스템

-개원 지역 데이터

여러 신도시 중에서 위례 신도시를 대상으로 하였다. 서울 지역의 데이터와 비교해야 하기 때문에 매출을 제외한 서울과 같은 데이터를 수집하였다. 위례가 서울, 성남, 하남 세 도시에 걸쳐져 있어 각 시에 해당하는 데이터를 수집하였고 성남과 하남의 경우 유동인구와 임대료에 대한 데이터가 없는 관계로 다른 지역과 비교하여 임의로 사용하였다.

> 인구, 연령별 인구 \_ 통계청

> 소득수준 \_ 국세통계포털

> 임대료 \_ 네이버부동산 참고

> 치과 의원 개수 \_ 보건의료빅데이터개방시스템

3.2 데이터 정합성 검사

분석을 위해 각 데이터의 정합성을 검사한 결과 결측값이 존재하는 데이터는 없었다.

3.3 데이터 전처리

데이터들을 분석하기 위해서는 모든 데이터들을 같은 형태로 만들 필요가 있었다. 마지막 매출을 예상하는 지역의 단위가 행정동이기 때문에 서울의 자치구 데이터를 세분화 시킬 필요가 있었다. 각 자치구의 값을 자치구마다의 행정동 개수로 나누어 행정동단위의 데이터로 처리될 수 있도록 한다.

연도별 데이터를 데이터프레임 형태로 정리하여 총 4개의 연도별 데이터 파일을 생성하였다. 또한 데이터를 수집할 때 각 데이터마다 조사 기간이 달랐기 때문에 한달을 기준으로 데이터들을 조정하였다. 연도별 데이터를 수집할 때 연도가 분기별로 나눠져 있는 데이터는 해당 연도의 4분기 데이터를 사용하였다. 데이터의 일관성을 위해 소수점이 나오는 데이터는 반올림하여 모두 정수의 형식으로 통일하였다.

서울 안에서도 자치구별로 편차가 큰 지역이 있어 특이값이 있는 행을 제외하고 분석을 진행할 수 있도록 하였다.

4. 분석 과정 및 결과

회귀 분석을 활용하여 개원 시 매출을 예측하려 한다. 서울 자치구 25개를 4개년으로 나누어 100개의 매출과 관련된 데이터를 회귀분석을 사용한다.

종속변수를 매출로 설정하고 독립변수(인구, 연령별 인구, 유동 인구, 생활 인구, 치과 의원 수, 소득수준, 버스 정류장 수, 임대료) 변화에 따른 모델을 구하기 위해 다중선형 회귀분석을 사용한다. 앞서 실행한 회귀분석의 적합성을 높이기 위한 변수 선택법으로는 단계적 선택법(stepwise selection)을 사용한다.

위와 같은 과정은 수집한 데이터들을 기반으로 R프로그램을 사용하여 진행되었다.

4.1 서울 매출 분석 결과

**[표1-1]**

|  |  |  |  |
| --- | --- | --- | --- |
| 독립 변수 | B | t-value | p-value |
| 상수항(Constant term) | 5.633e+07 | 6.208 | 5.5e-08 |
| 인구 | 2.268e+05 | 2.462 | 0.01669 |
| 20대 미만 | -2.279e+05 | -2.472 | 0.01631 |
| 20대 | -2.265e+05 | -2.458 | 0.01688 |
| 30대 | -2.275e+05 | -2.470 | 0.01636 |
| 40대 | -2.251e+05 | -2.446 | 0.01738 |
| 50대 | -2.273e+05 | -2.470 | 0.01636 |
| 60대 | -2.264e+05 | -2.455 | 0.01702 |
| 70대 이상 | -2.273e+05 | -2.467 | 0.01652 |
| 유동 인구 | -6.285e+01 | -2.845 | 0.00606 |
| 생활 인구 | -1.016e+02 | -3.349 | 0.00141 |
| 치과 의원 개수 | 1.711e+05 | 3.128 | 0.00272 |
| 소득 수준 | -1.182e+00 | -0.618 | 0.53868 |
| 버스 정류장 수 | 4.797e+04 | 2.913 | 0.00502 |
| 임대료 | -9.153e+01 | -1.142 | 0.25783 |
| **Adjusted R²** | 0.552 | | |
| **p-value** | 9.337e-09 | | |

\* B: 가중치, t-value: 추정치/표준 오차 p-value: 유의수준, Adjusted R²: 조정 결정 계수

[표1-1]의 결과를 단계적 선택법을 사용하여 변수를 다시 선택한다.

**[표1-2]**

|  |  |  |  |
| --- | --- | --- | --- |
| 독립 변수 | B | t-value | p-value |
| 상수항(Constant term) | 5.388e+07 | 6.638 | 9.66e-09 |
| 인구 | 2.388e+05 | 2.666 | 0.009819 |
| 20대 미만 | -2.400e+05 | -2.677 | 0.009537 |
| 20대 | -2.385e+05 | -2.660 | 0.009959 |
| 30대 | -2.396e+05 | -2.675 | 0.009581 |
| 40대 | -2.370e+05 | -2.648 | 0.010292 |
| 50대 | -2.393e+05 | -2.674 | 0.009618 |
| 60대 | -2.384e+05 | -2.656 | 0.010064 |
| 70대 이상 | -2.394e+05 | -2.671 | 0.009671 |
| 유동 인구 | -6.374e-01 | -2.906 | 0.005091 |
| 생활 인구 | -1.092e+02 | -3.959 | 0.000199 |
| 치과 의원 개수 | 1.836e+05 | 3.630 | 0.000582 |
| 버스 정류장 수 | 5.328e+04 | 3.814 | 0.000321 |
| 임대료 | -1.214e+02 | -1.910 | 0.060817 |
| **Adjusted R²** | 0.5565 | | |
| **p-value** | 3.744e-09 | | |

단계적 선택법을 사용한 후 독립변수 중 소득 수준이 삭제된다.

[표1-2] 모델의 회귀식은 다음과 같다.

서울 지역 예상 매출 = (5.388e+07) + (2.388e+05 \* 인구) - (2.400e+05 \* 20대 미만) - (2.385e+05 \* 20대) - (2.396e+05 \* 30대) - (2.370e+05 \* 40대) - (2.393e+05 \* 50대) - (2.384e+05 \* 60대) - (2.394e+05 \* 70대 이상) - (6.374e-01 \* 유동 인구) –

(1.092e+02 \* 생활 인구) + (1.836e+05 \* 치과 의원 수) + (5.328e+04 \* 버스 정류장 수) - (1.214e+02 \* 임대료)

Adjusted R²: 0.5565

p-value: 3.744e-09

4.2 지역 매출 분석 결과 및 시각화

4-2-1. 서울시 송파구 위례동 매출 분석 결과

**[표2-1]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **서울시 송파구 위례동** | | | | |
| 인구 | 20대 미만 | 20대 | 30대 | 50대 |
| 44,885 | 12,618 | 4,241 | 6,812 | 5,856 |
| 70대 이상 | 유동 인구 | 생활 인구 | 치과 의원 수 | 버스정류장 수 |
| 2,201 | 101,674 | 23,348 | 6 | 489 |
| **예상 월 매출** | | | | |
| 7,547,885 | | | | |

4-2-2. 성남시 수정구 위례동 매출 분석 결과

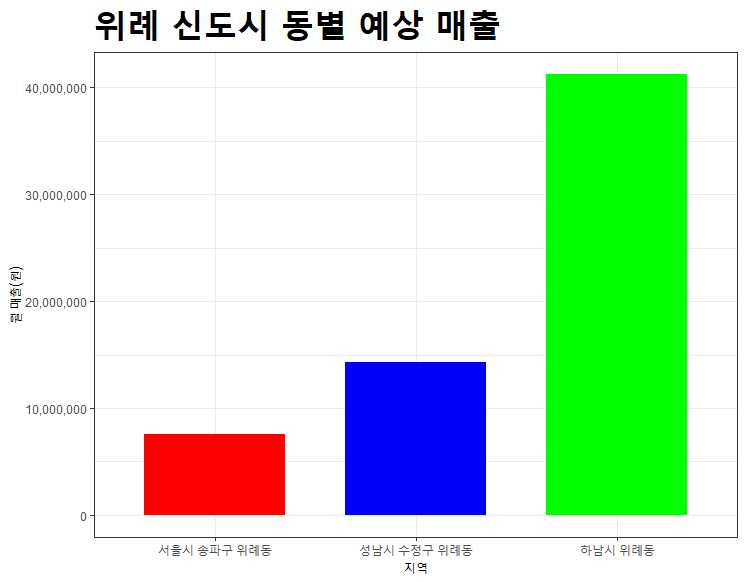
**[표2-2]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **성남시 수정구 위례동** | | | | |
| 인구 | 20대 미만 | 20대 | 30대 | 50대 |
| 45,498 | 11,456 | 4,664 | 7,231 | 6,758 |
| 70대 이상 | 유동 인구 | 생활 인구 | 치과 의원 수 | 버스정류장 수 |
| 3,434 | 69,598 | 20,000 | 29 | 433 |
| **예상 월 매출** | | | | |
| 14,280,555 | | | | |

4-2-3. 하남시 위례동 매출 분석 결과

**[표2-3]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **하남시 위례동** | | | | |
| 인구 | 20대 미만 | 20대 | 30대 | 50대 |
| 35,031 | 8,693 | 3,484 | 5,151 | 5,255 |
| 70대 이상 | 유동 인구 | 생활 인구 | 치과 의원 수 | 버스정류장 수 |
| 1,623 | 21,556 | 18,000 | 5 | 583 |
| **예상 월 매출** | | | | |
| 41,214,158 | | | | |



지역에 대한 회귀분석을 한 결과 조정 결정 계수(Adjusted R²) 값이 0.5565로 약 55%의 설명력을 가지고 있는 회귀식이 도출되었다. p-value의 값은 3.744e-09로 이 값이 0.05보다 작을 시 의미 있는 모델이라 할 수 있으므로 0.000000004 값이 나와 의미 있는 모델이라 할 수 있다.

매출의 마이너스 요소인 주요 변수로는 20대미만, 30대, 유동 인구가 있고 매출의 플러스 요소인 변수로는 총인구, 치과 의원 수, 버스 정류장 수가 주요하다고 나온다.

지역 데이터에 따라 예상 매출에 큰 차이가 존재하기 때문에 개원 시 주변 환경에 대한 고려는 필수라는 결론이 도출되었다.

5. 활용방안

지역 데이터들을 이용하여 예상 매출을 보여줌으로써 개원을 생각하고 있는 의사들에게 도움을 줄 수 있다.

본문에서는 치과만 대상으로 하였지만 다른 진료 과목의 데이터를 갖고 식을 새로 도출할 시 다른 과목의 매출도 예상할 수 있을 것이다.

6. 참고문헌