서울특별시 내 카페 입지 요건에 대한 분석

201902934 컴퓨터공학부 이호영

201703300 아랍어통번역학과 정현기

201901870 중국어통번역학과 서효민

한국은 ‘커피공화국’이라고 불릴 만큼 전 세계에서 손에 꼽힐 정도로 커피를 사랑하는 국가이다. 그 명성에 어울리게도, 서울에는 커피 전문점이 굉장히 빼곡하게 위치하고 있다. 언론사 ‘뉴스래빗’의 분석에 의하면 2020년 서울시의 프랜차이즈 커피 전문점은 총 1만 8535곳으로, 이는 2017년 대비 22%가 증가한 수치이다.**[[1]](#footnote-1)** 이러한 증가 추세로 인해 일반 소비자들은 서울 시내 어디든지 커피 전문점이 들어서 있을 것이라 유추할 수도 있으나, 그렇다고 이들에게 직접적으로 서비스를 제공하는 카페 점장들이 무턱대고 카페를 창업한 것이 아닐 것이다. 카페 입지를 한번 자리 잡으면 다시 바꾸기가 쉽지 않을 뿐 더러, 잘못된 입지로 인한 손해를 생각하면 카페 입지 선정에 큰 주의를 기울여야 할 것이다.

따라서, 본 보고서에서는 서울특별시 내 카페 점포수의 증가 현상에 어떠한 요인들이 작용하고 있는지 통계적으로 접근해보고, 나아가 본 연구를 통해서 카페 입점 시 고려해봐야 할 인적 · 지리적 요소들을 도출해볼 수 있을 것이다. 분석을 위한 프로그램으로는 ‘RStudio’ 를 활용하였으며, 서울 특별시 내 통계 자료는 ‘서울 열린 데이터 광장’과 ‘공공 데이터포털’ 두 웹사이트에서 수집할 수 있었다.

[R 모델링에 사용된 데이터]

|  |  |  |
| --- | --- | --- |
| **종속 변수** | 1 | 서울시 내 5개 프랜차이즈 카페 점포수 |

|  |  |  |
| --- | --- | --- |
| **독립 변수** | 1 | 평균생활인구 |
| 2 | 평균공시지가 |
| 3 | 상권변화지표 |
| 4 | 만족도지표 |
| 5 | 노령화지수 |
| 6 | 성비 |
| 7 | 지하철역 수 |
| 8 | 영화관 수 |

종속 변수로는 ‘서울시 내 자치구별 프랜차이즈 카페 점포 수’로 설정하였다. 프랜차이즈 카페에는 한국브랜드평판연구소가 선정한 2021년 카페 브랜드 평판 지수 TOP 5에 속한 ▲스타벅스 ▲투썸플레이스 ▲이디야 ▲빽다방 등 5개의 프랜차이즈 브랜드가 서울시에 보유한 점포 수를 기준으로 하였다.

첫 번째 독립변수로는 평균생활인구로 선정하였다. 생활인구란, 서울시와 KT가 공공데이터를 이용하여 추계한 서울의 특정 지역 및 특정 시점에 존재하는 모든 인구를 의미한다.**[[2]](#footnote-2)** 생활 인구 데이터로 서울에 거주하고 있지 않지만, 서울에서 활동하는 수도권 인구의 분포를 담고 있기 때문에 서울에서 활동하는 인구를 보다 면밀히 살펴볼 볼 수 있었다. 해당 연구에서는 2021년 한 해 동안 추계된 생활인구 데이터를 모델링으로 사용하기 위해 일 단위로 평균을 구했다.

두 번째 독립변수로는 평균공시지가로 선정하였다. 공시지가는 매년 토지소재지 자치구청장이 결정공시하는 1m2당 토지단가이다. 실제로 토지 단가가 카페 점포 수에 영향을 미치는 지 통계적으로 분석해보고자 이 변수를 독립변수로 선정하였다. 원본 데이터가 행정동 단위로 구분되어 있었기 때문에, 모델링에 사용하기 위해 구단위로 평균을 구했다.

세 번째 독립변수로는 상권변화지표로 선정하였다. 상권변화지표란 생존한 사업체의 평균 영업기간과 폐업한 사업체의 평균 영업기간을 4개 등급으로 나누어 상업공간의 지표를 책정하였다. “매우높음 – 높음 – 낮음 – 매우낮음”으로 등급이 구분되며, 등급이 높다는 것은 사업체의 영업 기간이 서울 평균 생존 영업기간보다 낮고, 서울시 평균 폐업영업기간보다 높은 상권이라는 것을 의미한다. 본 연구에서는 이 데이터를 모델링에 사용하기 위해 범주형 변수를 더미변수화 하였다.

네 번째 독립변수로는 만족도지표로 선정하였다. 만족도지표는 2021년 서울 시 2만가구주를 대상으로 실시한 서울 서베이 조사 중 가구주 조사 문항에서 발췌된 것으로, 서울 시민의 주거 및 교육 환경 만족도를 책정하기 위해 만들어진 데이터이다. 만족 정도는 5점 척도로 통계가 내어졌으나, ‘서울 열린 데이터 광장’ 측에서 이 자료의 점수를 10점 척도로 환산하였다.

다섯 번째 독립변수는 노령화지수로 선정하였다. 노령화 지수는 14세 이하 유소년 인구 100명 당 65세 이상 고령 인구의 비를 나타낸 것이다. 2021년 서울시 자치구별 노령화 지수를 모델링에 사용하였다.

여섯 번째 독립변수는 성비로 선정하였다. 성비는 여성 인구 100명당 남성 인구의 비를 나타낸 것이다. 2021년 서울시 자치구별 성비를 모델링에 사용하였다.

일곱 번째 독립변수와 여덟 번째 독립변수는 각각 자치구별 지하철의 수와 영화관의 수로 흔히 그 주변에 커피 전문점이 입점해 있을 것이라 추측되는 시설로 선정해보았다. 시점은 2021년 기준 데이터로 사용하였다.

Time-series Data 분석이 아닌 Cross-sectional Data 분석으로 접근하기 위하여 변수들의 시점을 2021년 기준으로 고정하였다.

[독립변수 분석 및 선별 과정]

독립변수를 선별하는 과정에서는 독립변수들 간 모든 부분집합을 모델링하여 돌리는 방법인 부분집합회귀 방법과 독립변수들간 상관관계 계수를 통하여 상관성이 높은 변수 위주의 선별 방법을 사용하였다. leap 패키지의 regsubset 함수로 모든 부분집합회귀를 이용하여 모든 부분집합 회귀를 통해 모델에 적합한 예측 변수를 도출하고자 했다. Regsubset 함수는 모든 X변수의 부분집합 중 모델링에 최적의 경우를 주는 함수이다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 평균생활인구 | 평균공시지가 | 지하철역수 | 상권변화지표 | 만족도지표 | 성비 | 영화관수 | 노령화지수 |
| 1 |  | ● |  |  |  |  |  |  |
| 2 | ● | ● |  |  |  |  |  |  |
| 3 | ● | ● |  |  |  |  | ● |  |
| 4 | ● | ● |  |  |  | ● | ● |  |
| 5 | ● | ● |  |  | ● | ● | ● |  |
| 6 | ● | ● |  |  | ● | ● | ● | ● |
| 7 | ● | ● | ● |  | ● | ● | ● | ● |
| 8 | ● | ● | ● | ● | ● | ● | ● | ● |

Regsubset 함수를 실행하고 나서는 종속변수와 독립변수 간의 상관성을 조사하였다. 상권변화지표를 제외한 모든 독립변수는 절댓값 3 이상의 상관계수가 나타났으나, 상권 변화지표의 상관계수는 0.01로, 종속변수와의 상관관계가 없는 수준으로 발견되어, 모델링에서 제외하였다.

또한, 변수가 많아질수록 유의성 검정에서 조합이 유의하지 않게되어, 적당한 변수들의 조합인 4~5개의 변수로 모델링을 시행하였다.

[모델 선정 및 분석]

위의 regsubset 함수 결과표와 상관관계 분석에 기반하여 변수 5개 모델, subset을 통한변수 4개 모델, 상관관계 분석을 통한 변수 4개 모델을 각각 mode1, model2, model3로 선정하였다.

• model1(상관관계 분석을 통한 변수 5개 모델링)

model1은 상관관계 분석을 기반하여 상관성이 높은 독립변수를 ‘평균생활인구, 평균공시지가, 지하철역수, 노령화지수, 영화관’ 등 5가지로 두고 모델링하였다. 그 결과, model1에는 두 가지의 문제점이 존재했다는 것을 발견할 수 있었다. 유의성 문제와 해석의 문제이다.

유의성 문제는 비교적 해결이 쉬운 문제였다. 독립변수 ’지하철역의 수’에서 유의성이 사라지는 결과가 나타났기에 해당 변수를 로그 변환해줌으로써 유의성을 부여해줄 수 있었다. 그러나, 심각한 문제는 그 이후에 나타난 해석의 문제였다. 상관 관계 플롯에 의하면 종속변수와 독립변수 ’지하철역의 수’의 상관계수는 +0.69 였으나, 양의 관계성을 보이던 것이 로그를 취함으로써 회귀계수가 음의 관계성으로 바뀌었다. 해석적 측면에서 이는 지하철 역이 감소할수록 카페수가 증가한다는 결정적 오류를 일으켰다.

• model2 (부분집합 이용한 변수 4개 모델링)

model2는 regsubset 함수 결과표에 기반하여 독립변수를 ‘평균생활인구, 평균공시지가, 성비, 영화관수’ 등 4가지로 두고 lm 함수로 모델링하였다. 모델의 유의성, 정규성, 등분산성, 잔차 이상치, 선형성 등 4가지 특성에는 큰 문제가 없었다.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -612.02320 94.75870 -6.459 2.68e-06 \*\*\*

log(평균생활인구) 39.41161 5.36445 7.347 4.23e-07 \*\*\*

평균공시지가 0.07310 0.01017 7.189 5.83e-07 \*\*\*

성비 1.20653 0.54808 2.201 0.03962 \*

영화관 3.93084 1.02775 3.825 0.00106 \*\*

■ 정규성 검정: P-value(0.95)>0.05로 귀무가설(H0) 입증 **▶ 정규성 검정 통과**

■ 등분산성 검정: P-value(0.95)>0.05로 귀무가설(H0) 입증 **▶ 등분산성 검정 통과**

■ 자기상관 검정: P-value(0.91)>0.05로 귀무가설(H0) 입증 **▶ 자기상관 검정 통과**

■ 이상치 테스트: P-value(0.24)<0.05로 귀무가설(H0) 기각 **▶ 이상치 테스트 통과**

모델에 큰 영향력을 행사하는 Influential Point로는 강남구의 데이터가 존재했다. model2를 기반으로 모델링 진행한 뒤 잔차 플롯과 QQ플롯 등의 그래프를 그려보니, 강남구의 데이터는 오른쪽 상단에 위치하여 값을 끌어당기는 플롯을 그리고 있었으며, 강한 레버리지를 일으켰다. 다른 데이터들에 비해 월등한 수치로 모델에 영향력을 행사하고 있었다. 따라서, 이 데이터를 제거한 채 모델링 다시 시도하여 보니, 독립변수 ‘성비’의 유의성이 사라지는 결과로 나타났다.

• model3 (상관관계 분석을 이용한 변수 4개 모델)

model3은 상관관계 분석을 기반하여 독립변수를 ‘평균생활인구, 평균공시지가, 노령화지수, 영화관 변수 4개를 이용하였다.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

log(평균생활인구) 3.859 0.81354 4.744 0.000124 \*\*\*

평균공시지가 0.054 0.01428 3.824 0.0010060 \*\*

노령화지수 -0.1743 0.04953 -3.520 0.002153 \*\*

영화관 3.26870 1.25523 2.604 0.0016977 \*

■ 정규성 검정: P-value(0.2245)>0.05로 귀무가설(H0) 입증 **▶ 정규성 검정 통과**

■ 등분산성 검정: P-value(0.33)>0.05로 귀무가설(H0) 입증 **▶ 등분산성 검정 통과**

■ 자기상관 검정: P-value(0.76)>0.05로 귀무가설(H0) 입증 **▶ 자기상관 검정 통과**

■ 이상치 테스트: P-value(0.4)<0.05로 귀무가설(H0) 기각 **▶ 이상치 테스트 통과**

평균생활인구는 회귀계수를 조정하기 위함과 동시에, 해석적인 측면에서 수월함을 가져가지 위하여 log변환을 하였다. 각 정규성, 등분산성, 자기상관, 이상치 검정은 r라이브러리를 이용하여 귀무가설을 입증하거나 기각하는 방식으로 1차적으로 검정을 진행하였으며, gvlma 패키지를 이용하여 선형성 및 종속변인 연속형 변수 여부, 이분산성에 관한 검정을 한 번 더 진행하여 이중 체크를 진행하였다. 수치적인, 검정을 진행한 후, 플롯을 그려서 앞서 수치적으로 검정했던 부분들을 한 번 더 확인해볼 수 있었는데, 역시나 플롯 상 드러나는 Influential Data인 강남역을 제거하였다. Model2와 달리 Model3은 Influential Data를 제거하여도, 유의성에서 문제가 발생하지 않았으며, 각종 플롯이나 검정 수치들이 이상 없음을 확인하였다. 자세한 수치나 플롯 그림은 발표ppt에 기재되었다.

• 결론

위 모델 선정 과정에 기반하여, 모델 3번을 최종 모형으로 선정되었다. 이 모형에 의하면, 다른 변수가 동일할 때 평균생활 인구가 1% 증가할수록 카페 점포 수는 3.86개 증가하였다. 다른 변수가 동일할 때 평균공시지가가 한 단위 증가 즉, 만원이 증가하면 카페 점포 수는 0.05개가 증가하였다. 다른 변수가 동일할 때 영화관이 한 개 증가하면, 카페 점포수는 3.27개가 증가하다. 이러한 양상과는 반대로 노령화지수가 1 감소할 때 카페 점포 수는 0.17개가 감소하였다.

결과를 유추해보자면, 해당하는 변수 외 다른 변수가 동일할 때, 평균생활인구는 많을수록, 평균공시지가가 높을수록, 구역 내 영화관의 개수가 많을수록 카페의 점포 수가 증가한다는 결과가 나타났다. 이와는 반대로, 구역내 인구의 나이가 많을수록 즉, 노령화의 수치가 클수록 카페 점포수는 감소한다는 결과를 얻었다.

그러나 모델링 과정과 결과를 검토하는 과정에서 세 가지의 한계점이 발견되었다.

첫째, 모델의 신빙성이다. 모델링에 사용된 ‘서울 내 자치구별 카페 지점 수’라는 데이터는 오로지 대형 프랜차이즈 커피 전문점을 대상으로 집계한 데이터로, 구역내 개인 카페의 수는 고려하지 않은 수치였다. 따라서, 통계상 대형 프랜차이즈 지점을 벗어나 개인카페의 영역까지 확장한다면 본 모델링의 신빙성이 다소 떨어졌다.

둘째, 시간의 경과이다. 모델에 사용된 데이터들의 통계 시점은 모두 2021년이었다. 그렇기에 2022년 현재 시점에서는 해당 모델에 사소한 오차가 발생할 수 있을 것이다.

셋째, 해당 모델에 영화관 이외에도 지하철과 같은 여타 시설의 데이터를 독립 변수가 추가되지 않았다는 점이다. 이 한계점에 기반하여 원래 기획했던 연구 방안에 지하철이라는 독립 변수를 추가하였으나, 위의 과정에서 보았듯 해석상의 측면 오류에 직면하다. 만약, 적합한 데이터를 획득 시 모델링의 결과가 달랐을 가능성도 충분히 존재했다.

대부분의 한계점이 존재한 근본적 이유는 원하는 데이터셋의 부재였다. 공공데이터를 활용하여 원하는 데이터로 가공하는데도 한계가 존재하였다. 따라서, 데이터셋을 좀 더 구할 수 있다면, 프렌차이즈 같은 카페들이 아닌 소형 카페를 포함하여 집계한 데이터 조사, 모델링에 사용된 데이터의 최신화 및 Cross-sectional Data분석이 아닌, Time-series Data 분석 같은 다른 계열의 분석, 다른 시설의 데이터에 기반한 모델링 재구축 등 세 가지의 사항에 기반하여 연구를 지속해 보고자 한다.

1. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)