1. 논문제목 (국문 혹은 영문)

(국문) - 빅데이터 분석을 통한 서울시 골목상권 분석

(영문) - Analysis of market district hinterland in Seoul by ‘Big Data’ analysis

2. 분야 : 경영정보

3. 저자명 및 소속(국문/영문) :

오흥록/HeungRok Oh: 수학사, 중앙대학교, 더 조은 IT 아카데미
배범석/BumSeok Bae: 컴퓨터정보통신공학사, 홍익대학교, 더 조은 IT 아카데미
(교신)문혜정/HyeJung Moon: 정책학박사, 컴퓨터공학석사, 서울과학기술대학교 겸임교수, 아이엘피 대표, 윌비솔루션 이사

[E-mail]
오흥록: heungrokoh@gmail.com
배범석: qoqjatjr10@naver.com
(교신)문혜정: hyejung.moon@gmail.com

빅데이터 분석을 통한 서울시 골목상권 분석

오흥록*, 배범석**, 문혜정***

초 록

서울시는 빅데이터 프로젝트의 목적으로 중소형 사업 중심의 골목상권 데이터와 분석 결과를 공개하고 있다. 서울시의 데이터는 다양하고 풍부하지만, 대부분이 통계수치의 제시로 구성되어 있다. 또한, 서울시에서 제공하고 있는 '신규창업위험도' 지표는 개·폐업률 중심의 지표로써 업종 및 상권별 특징을 설명하지 못하고 있다. 이러한 점에 착안하여, 본 연구자는 서울시의 풍부한 데이터를 활용하여 업종별로 성격이 비슷한 지역구로 구성된 상권을 정의하고 대표적인 특징을 파악하고자 한다. 따라서 연구의 대상은 서울시에서 제공하는 2015년부터 2016년까지의 공개데이터로써 서울시에서 지정한 1008개의 골목상권과 상권배후지이다. 주제영역은 아파트, 집객시설, 소득소비, 업종별매출, 인구(상주, 유동, 직장인 등), 점포 개·폐업률 등이다. 연구절차는 1. 서울시 공공데이터를 수집 2. MySQL로 통합데이터베이스를 구축 3. MsExcel을 이용하여 기술분석 수행 4. R 프로그래밍 언어를 활용하여 상관분석과 회귀분석 수행 5. Tableau 툴을 통한 시각화이다.

분석결과 월매출액과 상관성이 있는 변수는 아파트평수, 인구(상주, 유동, 직장인 등), 요일별 매출과 연령별 매출, 지출비용 등 이다. 업종으로 구분된 각 구별 상관계수를 가지고 군집분석을 통해 서울시 25개 구를 5개 군집으로 나타냈다. 군집은 주로 인접한 지역구가 함께 구성되었다. 분석결과 강남일대 군집의 매출은 고가보다는 저가 아파트 단지 수나 비아파트 세대 수와 상관관계가 높았다. 반면에 도봉구 중심의 군집은 연령별 매출 상관계수의 편차가 다른 군집에 비해 작고 특징이 골고루 나타나는 특징을 볼 수 있었다. 이를 통해 업종에 따라 다르게 형성되는 서울시 구 단위 군집의 구성요소를 확인하고 각 군집별 특징 및 매출상관요인을 확인할 수 있었다.

주제어: 빅데이터, 상관 분석, 군집 분석, 골목상권, 상권배후지

* heungrokoh@gmail.com 수학사, 중앙대학교, 더조은IT아카데미
** qoqjatjr10@naver.com, 컴퓨터정보통신공학사, 홍익대학교, 더조은IT아카데미
*** 교신저자, hyejung.moon@gmail.com 정책학박사, 컴퓨터공학석사, 서울과학기술대학교 겸임교수, 아이엘피 대표, 윌비솔루션 이사

Ⅰ. 서론

공개 데이터를 이용한 빅데이터 분석이 상권영역에서도 활발하게 이루어지고 있다. 서울시에서는 골목상권서비스를 통해 골목상권 중심의 정보를 제공하고, 중소기업청은 상권정보시스템을 통해 발달상권중심의 정보를 제공하고 있다. 이러한 정보는 매우 다양하고 상세하게 제공되지만, 단순한 통계적 수치의 제시에 그치는 경우가 대부분이었다.

본 연구는 이러한 통계적 수치들을 활용하여 업종별로 매출과 상관있는 요소에는 어떤 것이 있으며, 비슷한 성격을 가지는 지역의 존재와 지리적으로 어떻게 위치해 있는지 확인하고자 시작하게 되었다. 상권 데이터 중 서울시에서 공개적으로 제공하는 골목상권과 상권배후지 데이터를 이용, 분석하여 이를 확인하고자 하였다. 이를 통해 업종별로 성격이 비슷한 지역구로 구성된 상권을 정의하고 대표적인 특징과 지리적 위치관계를 파악하고자 한다.

Ⅱ. 이론 및 선행문헌검토

2.1 이론

상관분석(Correlation Analysis)은 두 변수가 어떠한 관계에 있는지 파악하는 통계적 분석방법이다. 상관분석은 두 변수간의 인과관계를 나타내는 것이 아니라 관계의 추세를 나타내는 것이다. 상관분석에서 사용하는 상관계수로는 대표적으로 피어슨 상관계수(Pearson Correlation Coefficient), 스피어만 상관계수(Spearman's Rank Correlation Coefficient), 켄달의 순위상관계수(Kendal's Rank Correlation Coefficient)가 있다. 그 중 피어슨 상관계수는 연속형 변수에 사용된다. 상관계수(r) = X와 Y가 함께 변하는 정도 / X와 Y가 따로 변하는 정도를 의미한다. X와 Y가 동일방향이면 +1, 전혀 다르면 0, 반대방향이면 -1값을 가진다. 결정계수는 R^2 으로 계산하며 이것을 X로부터 Y를 예측할 수 있는 정도를 의미한다.

회귀분석(Regression Analysis)은 종속변수가 독립변수로부터 어떤 영향을 받는지 그 인과관계를 파악하는 분석법으로, 일정한 패턴을 찾아내어 무엇인가를 예측할 때 사용할 수 있다. 회귀분석은 다음과 같은 표준 가정을 필요로 한다. 1. 오차항(residuals)의 등분산성, 2. 오차항의 평균은 0 3. 데이터의 분산은 정규분포 4. 독립변수 사이에 상관관계가 없어야 한다. 5. 시간에 따라 수집한 데이터들은 잡음의 영향을 받지 않아야함. 특히, 독립변수들 간에 상관관계가 있는 경우를 다중공선성(Multicollinearity)의 문제라고 한다.

주성분분석(Principal Component Analysis)은 Pearson이 창안한 수학적 기법으로 상관관계가 있는 변수들을 결합해 상관관계가 없는 변수로 분산을 극대화하는 변수를 추출하는 방

법이다(K.Pearson, 1901). 주로 변수가 많은 분석에서 변수들 간에 내재하는 상관관계를 이용해 소수의 주성분으로 차원을 축소하는데 이용되는 방법이다. 대표적으로 다중회귀분석에서 다중공선성(Multicollinearity)의 문제를 해결하기 위해 사용한다.

클러스터 분석(Cluster analysis)이란 주어진 데이터들의 특성을 고려해 데이터 집단(클러스터)을 정의하고 데이터 집단의 대표할 수 있는 대표점을 찾는 것으로 데이터 마이닝의 한 방법이다. 클러스터란 비슷한 특성을 가진 데이터들의 집단이다. 반대로 데이터의 특성이 다르면 다른 클러스터에 속해야 한다.

위 이론을 기반으로 서울시 공공데이터를 분석하여 독립변수간의 상관관계와 지역별 특성을 연구하고자 했다.

2.2 선행문헌검토

이경주 외(2015)는 공간가중회귀모형을 통해 강원지역 상권 현황을 진단하고 활성화를 위한 정책적 시사점을 논의하였다. 분석결과 강원도 내에서도 각 독립변수의 영향력이 다르게 나타났다. 따라서 상권별로 상이한 영향력을 가지는 변수들을 고려한 차별적 정책방안이 필요하다는 결론을 내렸다.

구자용(2015)은 서울시 지역별 트윗 데이터 내용의 분포를 분석하였다. 데이터를 지도화한 결과, 트윗 데이터의 분포는 유동인구가 많은 지역과 비례하며, 도로와 인접한 지역에서 주로 발생하였다. 따라서 공간정보 빅데이터의 분석을 통해 그 지역에서 나타나는 현상, 특성을 파악할 수 있다는 점을 확인하였다.

정대석, 김형보(2014)는 경기도 31개 시군별 업종별 업소 수를 인구 원단위를 통해 비례상관관계를 차이가 존재하는지 분석하였다. 매출 영향 요인에 대한 회귀분석을 통해 음식, 생활서비스, 스포츠 업의 매출에 가장 큰 영향을 미치는 요인은 해당 지역의 공동주택 기준시가이다. 하지만 소매 및 관광 업종은 다른 요인이 존재하여 일반적인 매출 설명이 어려운 업종이었다. 따라서 도시계획이나 개발계획 상의 용도별 면적을 계획할 때, 해당 지역의 특성 파악이 우선시 되어야 하고 세부 용도별 용지 면적 설정 시 업종별 상관성을 고려해야 한다는 결론을 내렸다.

이명호(2016)는 서울시 내 지역 상권에 따라 발생하는 매출 영향력을 공간가중회귀모형을 통해 분석하였다. 종속변수로는 매출액이 선택되었으며 독립변수로는 주거인구, 유동인구, 직장인구, 평균소득, 지역상권 내 상점의 수이다. 연구는 서울시를 19,440개의 소지역으로 구분하고 매출에 미치는 개별 독립변수의 영향력을 다른 변수들이 일정하다는 가정 하에 각각 회귀계수를 비교하였다. 분석결과 매출액에 미치는 변수는 지역에 따라 상이하게 나타났으며 선정된 변수 외에도 제 3의 변수가 존재한다는 결론을 얻었으며, 다양한 변수와 분석방법을 고려하지 못한 한계점을 가졌다.

Ⅲ. 연구설계

3.1 연구 질문

서울시에서 제공하는 '우리마을가게 상권분석' 서비스에서 주어진 '빅 데이터'를 기존의 통계분석이 아닌 빅데이터 분석을 통해서 해당 '우리마을가게 상권분석' 서비스에서 제공되지 않는 새로운 통찰을 확인해 보자는 취지에서 연구를 진행했다.
골목상권의 발달업종을 파악하고 그 중 대표업종을 선택해 월 매출액을 기준으로 서울시 25개 구의 군집특징에 대해 통찰을 시도 했다. 또한, 세부적으로는 군집을 구성하는 지역구별 특징을 파악하여 군집의 통찰뿐만 아니라 지역구별 통찰도 연구진행했다.
구체적으로 본 연구에서 진행하고자 하는 질문은 다음과 같다.

Q1. 서울시 골목상권과 상권배후지의 업종별 지역구별 매출에 미치는 요인은 무엇인가?
Q2. 서울시 골목상권과 상권배후지의 업종별 매출에 미치는 요인이 유사한 지역구는 무엇인가?
Q3. 서울시 골목상권과 상권배후지의 업종별 유사한 지역구들의 특징은 무엇인가?

3.2 연구방법

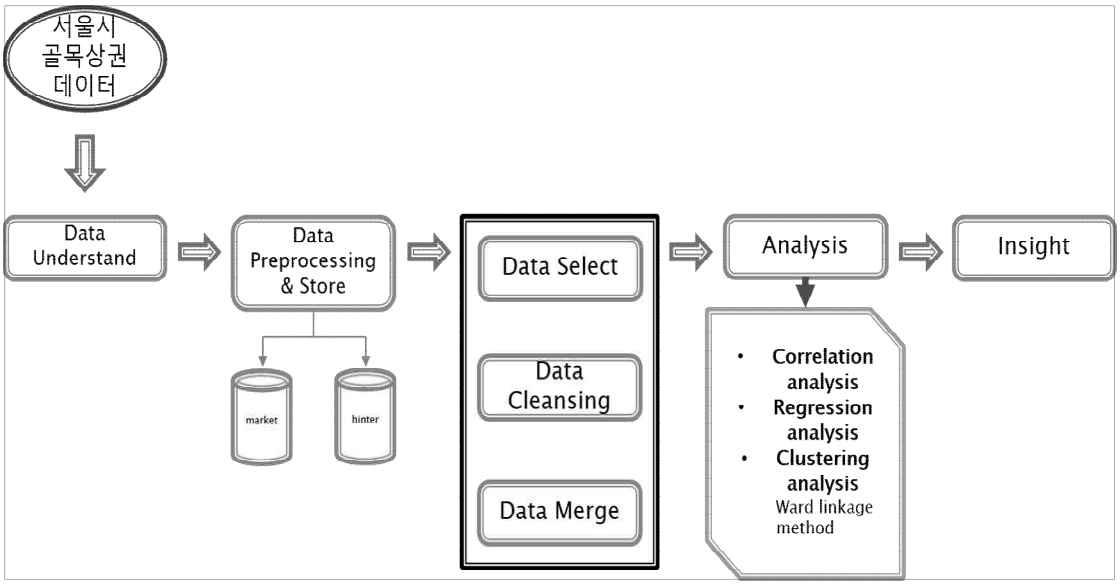
서울시에서 제공받은 상권분석 관련 빅데이터는 2013~2016년 동안의 거주정보, 집객시설, 유동인구, 거주인구, 직장인구, 지출비용, 매출정보, 업종정보, 구 및 상권 코드, 업종코드 등으로 구성된 골목상권과 상권배후지 정보이다.

해당 정보들을 파악하고 데이터를 관찰한 결과, 데이터의 수집시기와 갱신주기가 서로 달라서 정보가 지속수집 된 2015~2016년 정보를 분석의 대상으로 지정하였다.

데이터의 저장은 골목상권 8개의 테이블과 상권배후지 8개의 테이블을 각각의 MySQL 서버에 저장하였다. 데이터 정보에 관한 테이블은 상권배후지가 속해있는 MySQL서버에 저장했다. 따라서 두 개의 MySQL 서버에 각각 골목상권 8개의 테이블과 상권배후지 9개의 테이블을 저장하여, 클라이언트가 골목상권과 상권배후지 DB서버에 접속 가능하게 구성했다.

필요한 데이터를 골목상권과 상권배후지 영역으로 나누어 R studio를 통해 불러들이고 선택, 가공 또는 병합함으로써 분석수행을 위한 통합테이블을 만들었다.

분석은 통계적 분석방법을 기반으로 한 귀납적 방법으로써, 상관분석, 회귀분석, 주성분분석, 군집분석 방법을 선택하였고 도구는 R studio를 사용하였다. 분석 결과는 R studio와 MS Excel을 사용하여 살펴보았으며 시각화는 Tableau 프로그램을 사용하였다.



<그림 1> 연구 절차

3.3 자료의 수집

본 연구의 자료는 서울시 '우리마을가게 상권분석' 서비스에서 사용하고 있는 데이터를 기반으로 연구를 진행 하였다.
아래는 '우리마을가게 상권분석' 서비스에서 제공하는 데이터 출처 이다.

데이터 명	주요내용	갱신주기	데이터 출처	제공 지역
1.상가(업소 정보	업종별 업소정보(업종, 주소, 전화번호)	분기	서울시	서울시 전체
2.인허가 업소 정보	- 인허가 대상 업종 업소 정보 - 상권별 개폐업률 산출 - 상가 이력정보	월	서울시	서울시 전체
3.사업자 등록(휴폐업) 정보	- 휴, 폐업 신고 사업자 정보(상가/업소_기업DB 보강)	분기	서울시	서울시 전체
4.임대시세	- 상가 임대시세 조사 지표 - 행정중별, 자치구별 상가 임대시세 정보	분기	한국감정원	서울시 전체
5.골목상권영역	- 서울시 발달상권 1,000여 개 영역 - 골목상권 단위 정보	년	서울시	서울시 전체
6.발달상권영역	- 서울시 주요 발달상권 260개 영역 - 발달상권 단위 정보	년	소상공인시장진흥공단	서울시 전체
7.매출/소비정보(BC카드)	- 신용카드 매출정보 - 골목별, 상권별 매출정보(매출액, 거래건수) - 성/연령대별, 시간대별, 요일별 거래패턴 정보	월	BC카드	서울시 전체
8.매출/소비정보(KB카드)	- 신용카드 매출정보 - 골목별, 상권별 매출정보(매출액, 거래건수) - 성/연령대별, 시간대별, 요일별 거래패턴 정보	월	KB카드	서울시 전체
9.카드비중추정 정보	- 지역별 업종별 신용카드사별 점유비 - 신용카드 정보와 융합하여 추정매출 산출	년	나이스지니데이터	서울시 전체
10.원금비중추정 정보	- 지역별 업종별 현금영수증 및 현금 비중 - 신용카드 정보와 융합하여 추정매출 산출	년	나이스지니데이터	서울시 전체
11.집단의 추정 유동인구 정보	- 성, 연령, 요일, 시간대별 유동추정 인구 - 통신사 데이터, 정책요인, 고룡카드 데이터를 융합하여 도로 단위의 유동인구 추정 정보	월	서울시	서울시 전체
12.유동인구	- 이동통신 통화량기반 유동인구 정보(50x50 Cell)	월	KT	서울시 전체
13.교통카드 정보	- 지하철 및 버스정류장별 승하차, 유동인구수	월	서울시	서울시 전체
14.사업체 통계 DB	- 집계구별 업종별 사업체통계	년	통계청	서울시 전체
15.사업체 조사 DB	- 사업체 총 조사 데이터	년	서울시	서울시 전체
16.주거인구	- 행정구역별 주민등록 통계 데이터를 건물단위별 가구수 및 성별/연령대별 인구수 추정	반기	행정자치부	서울시 전체
17.직장인구	- 50m Cell 단위의 성/연령별 직장인구 정보	반기	나이스지니데이터	서울시 전체
18.직업직종	- 행정구역 단위 업종별 종사자 통계 정보	년	통계청	서울시 전체

19.아파트 DB	- 아파트 단지/동 단위 가구수 정보 - 면적별/기준시기별 가구수 정보	년	서울시	서울시 전체
20.소득데이터	- 블록단위로 가공된 성별 / 연령대별 10분위 기준 소득추정액 정보	반기	나이스지니데이터	서울시 전체
21.소비특성 데이터	- 블록별 소비유형별 비율(식료품, 의류 및 신발,가사용품, 의료, 할것, 여가, 문화, 교육, 커피, 주류)	반기	나이스지니데이터	서울시 전체
22.블록데이터(블록영역)	- 서울시 6만8천 여개의 블록영역	년	오픈데이트	서울시 전체
23.블록유형화(블록속성 데이터)	- 서울시 6만8천 여개의 블록에 대한 배후지 속성 정보	년	오픈데이트	서울시 전체
24.건물 DB	- 건물의 용도, 층수, 면적, 주출입구, 건축일자, 주차장여부, 접도정보, 토지의 형상 등(사주소건물+건축물대장)	년	서울시	서울시 전체
25.버스정류장	- 시내 버스 정류장 정보(버스노선, 위치정보)	월	서울시	서울시 전체
26.지하철역	- 지하철역사 정보(노선, 위치정보)	년	서울시	서울시 전체
27.주요/집객 시설	- 주요/집객시설 구분 및 위치정보(관공서, 금융기관, 병원, 학교, 유통점, 문화관광 / 영화관, 숙박시설, 교통관련 시설)	년	각급기관	서울시 전체
28.도로명 주소 도로링크	- 도로명 주소 지도 데이터-도로구간 데이터	반기	서울시	서울시 전체

업종분류(상점밀착형 43개 업종)

업종	세분업종				
외식업(10)	한식음식점	중국음식점	일식/수산물	양식음식점	분식음식점
	제과점	패스트푸드점	치킨전문점	호프/간이주점	커피음료
서비스업(22)	입시/보습학원	외국어학원	예체능학원	일반의원	치과의원
	한의원	부동산중개업	인테리어	노래방	PC방
	보육시설	노인요양시설	헬스클럽	당구장	골프연습장
	미용실	피부관리실	네일숍	여관업	세탁소
	자동차수리	자동차미용			
도소매업(11)	슈퍼마켓	편의점	컴퓨터판매/수리	호대폰	정육점
	과일채소	의류점	화장품점화	약국	문구점
	화장품				

<그림 2> 서울시 ‘우리마을가게 상권분석’ 데이터 출처

‘우리마을가게 상권분석’에서 사용한 데이터 중 ‘서울 열린데이터 광장(http://data.seoul.go.kr/)’을 통해 얻을 수 있는 공개데이터만을 수집하였다.

<그림 3> 홍대 앞 골목상권 영역

<그림 4> 골목상권과 배후상권

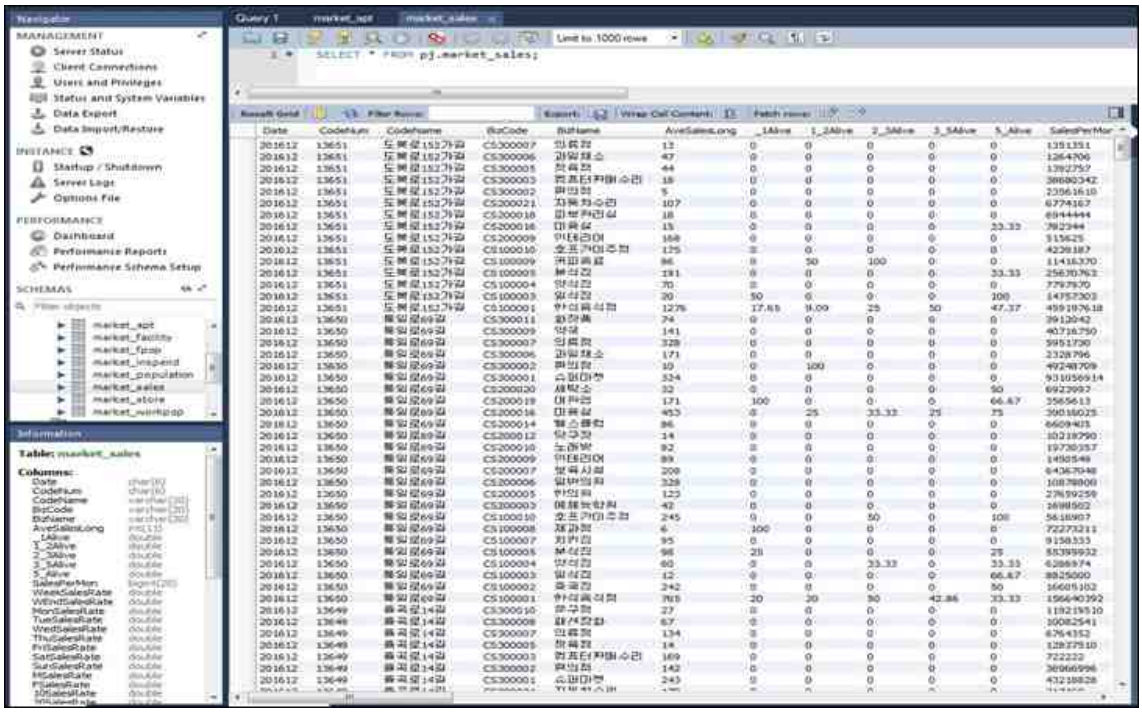
3.4 데이터 선택, 전처리, 병합

데이터 선택, 전처리, 병합과정에서는 MySQL에 저장된 데이터를 R 프로그래밍에서 'RMySQL' 패키지를 사용해 불러들이고 이를 분석에 사용했다.

테이블은 추정유동인구, 직장인구, 상주인구, 추정매출, 소득소비, 점포, 아파트, 집객시설이 골목상권과 상권배후지로 나뉘어 각각 존재하였고 상권지수지표와 상권영역설명 데이터를 포함한 총 18개로 구성되었다. 이 중 상권지수지표는 데이터기간이 분석하고자 하는 기간에 미치지 못하여 제외하였다. 그리고 나머지 데이터의 데이터기간에 다소 차이가 있어 2015년부터 2016년까지의 기간을 연구대상으로 선택하였다. 데이터 갱신주기 역시 데이터마다 상이하였으나 보다 세분화된 분석을 위해 갱신 최소 단위인 월 단위로 통일하였다.

데이터는 총 25개 구와 46개 업종, 1008개 골목상권, 120개의 변수로써 24개월간의 축적된 데이터이다. 46개의 업종 중 서비스업전체, 외식업전체 등 포괄적인 4개의 업종과 결측치가 많은 노인요양시설을 제외한 41개의 업종에 대해서만 분석을 진행하였다. 또한, 120개의 변수 중 ‘점포 수’는 점포테이블과 추정매출테이블에서 각각 다른 값을 가졌기 때문에 제외하여 119개의 변수만을 선택하였다. 이 중 분석에 사용된 변수는 고유번호, 날짜 등의 5개 변수를 제외한 114개로써, 월매출액과의 관계를 파악하는 것이 목적이기 때문에 월매출액과 나머지로 구분하였다. 결측치를 갖는 변수들은 전체에 비해 양이 적고 결측치가 아닌 변수와 비교불가하기에 분석의 대상에서 제외하였다.

정제된 데이터들을 분석모델에 적합한 형태로 만들기 위해, R studio를 사용하여 골목상권과 상권배후지 각 8개의 테이블을 병합하였다. 그 결과 골목상권은 446,385개, 상권배후지는 718,362개의 데이터로 구성되어 각각 114개의 변수를 갖는 테이블을 만들었다.



<그림 5> MySQL에 저장된 테이블 형태

3.5 분석방법

1) 상관 분석

상관관계는 한 변수의 값이 변화함에 따라 다른 변수의 값이 증가하거나 감소하는 관계의 추세를 나타내는 지표이다. 즉, 두 확률변수 사이의 관련성을 파악하는 방법이다. 그래프의 기울기에 따라 양의 상관관계 음의 상관관계로 나눌 수 있는데 비슷한 기울기를 보여도 변수들의 밀도의 차이를 보여주는데 이 밀도를 표현한 숫자를 상관계수라고 부른다. 상관계수를 이용하면 절대적인 비교는 불가하지만 상대적인 비교는 가능하다. 따라서 변수들 사이의 상관계수를 비교하여 상대적으로 상관성이 높은지 낮은지(양 또는 음)를 판별할 것이다.

업종별, 지역구별로 월매출액과 나머지 변수들의 상관관계를 R프로그래밍 언어의 상관분석 도구를 이용하여 도출한다. 상관분석법은 분석대상 변수들이 금액, 인구수와 같은 연속형 변수이기 때문에 피어슨 상관계수를 사용하였다. 상관분석의 대상은 수익에서 큰 영향을

미치는 월 매출액과 나머지 변수들 즉, 아파트(면적·가격별 세대 수), 유동인구(나이, 시간대, 성별, 요일), 주거인구(나이, 성별), 직장인구(나이, 성별), 매출(요일, 성별, 나이), 수입 및 지출(식품, 교육 등) 등 113개의 변수를 선택하였다. 업종 및 지역구별 비교를 위해 41개 업종별, 25개 지역구별로 나누어 각각에 대해 상관분석을 하였다.

```
mer = merge(mer,inspend_data[-3])
mer = merge(mer,population_data[-3])
sales_data_ = sales_data[,-c(3,6:11,13:29,47)]
mer = merge(mer,sales_data_)
store_data_ = store_data[,-c(6:11)]
mer = merge(mer,store_data_)
mer = merge(mer,workpop_data[-3])
mer = merge(mer,biz)
mer = merge(mer,Gu_info)

colnames(mer)
str(mer)
#상관 돌리기
cor_gu_biz = data.frame(1:1)

for(i in gu$GuCode){ # gu 분류
  gu_biz_mer = mer[mer$GuCode==i,]
  temp_code = order(unique(gu_biz_mer$BizCode))
  temp_code = unique(gu_biz_mer$BizCode)[temp_code]
  print(i)
  for(j in temp_code){ # biz 분류
    biz_mer_be = gu_biz_mer[gu_biz_mer$BizCode==j,]
    biz_mer = biz_mer_be[,-c(1:5,97,119)]
    sum_co = as.data.frame(corr.test(biz_mer)[1])[,74] # SalesPerMon만 추출
    cor_gu_biz = cbind(cor_gu_biz,sum_co)
    print(j)
  }
}
```

<그림 6> R 프로그래밍 언어를 활용한 25개 구 & 41개 업종 상관관계 분석코드

APTAvePrice	TotalFPop	MFPop	WFPop	X10FPop	X20FPop	X30FPop	X40FPop	X50FPop	X60FPop
-0.559820068791568	0.649144335835411	0.618971098815173	0.677665459416342	0.777461606269647	0.73507178450011	0.649253300818444	0.646974739469852	0.550061646710131	0.5706914366
-0.360202282430383	0.320827980629 75	0.294168930142327	0.348480090896547	0.585977470019313	0.423812815993042	0.362978615387753	0.350983739345529	0.189938920713433	0.1906466910
-0.38887170719453	0.279533868351382	0.276122900977675	0.280272767273565	0.358438561378275	0.324383423879907	0.309316172022864	0.296222623684327	0.183381595365505	0.1841177033
-0.417122517518544	0.274709372564258	0.253808703495338	0.297304181222753	0.382077364476162	0.360380913826611	0.287251597541229	0.269055749256881	0.177236518446192	0.1890760666
-0.519672860076176	0.361977698014817	0.338542958181217	0.386521526517031	0.489924044888912	0.446235565 31924	0.357883356344333	0.347269113570507	0.266533752489685	0.2888195438
0.125697161181261	-0.123313223257509	-0.135081632869522	-0.106886766597798	0.141341395790346	0.027719085591835	-0.06462558053434 03	-0.128846909909309	-0.259689453842357	-0.2665579740
0.229402565362736	0.009624939261155 79	0.006191382147810 87	0.013805144741100 5	0.162566764131369	0.116641699936117	0.058468828471953 4	0.009090251011874 1	-0.097071696665794 6	-0.1137595920
-0.351973451950982	0.246067822871998	0.225557321629631	0.268091795047139	0.280235312538772	0.355352623162147	0.227484567076679	0.200509141290503	0.180148902589422	0.2259867766
-0.325032818837142	-0.007722619484718 36	0.000574789689667 238	-0.018207995180630 4	0.118630389869805	-0.051391267725672 9	0.014671976762909 2	0.033042667632030 2	-0.025534702517820 4	-0.0674996535
-0.374510505372736	0.052886883371266 4	0.019083089387367 4	0.094149031609496 8	0.381327599099872	0.197748169200659	0.038538765208191 3	0.017596089842653 8	-0.050261029704231 2	-0.0191363537
-0.111233023792451	0.145992224784 62	0.125744443473118	0.169409022980125	0.311679763981605	0.286159167443875	0.169535184660977	0.118920817794253	0.026658756015691	0.0485284492
-0.350503048386662	0.278446127057053	0.241412854036747	0.321316151848 84	0.520767162250757	0.456499856881166	0.292080101361527	0.239411688231 58	0.138886967158988	0.1433228082
-0.661895246281417	0.833338835400779	0.809028560290892	0.853334237009671	0.734093983658408	0.821043459852301	0.762113678207826	0.800055264115425	0.838325235657233	0.8738349955
-0.314868475971207	0.283853858507502	0.285554782628818	0.277847693758 66	0.199147724006837	0.324649480231108	0.256588662266574	0.259057563290966	0.276150397957229	0.2726376800
0.529495294234339	-0.51271801530761	-0.502715507788428	-0.522478215103387	-0.392540715005305	-0.521919473578714	-0.509236730394082	-0.496702099609849	-0.490539628495012	-0.5427510344
0.047529087894 9584	-0.207694560538259	-0.175206533436363	-0.245438860080628	-0.245571233762856	-0.241590610359538	-0.183781168551194	-0.173529477663427	-0.203173426602022	-0.2068270986
-0.334915168012087	0.415725624371665	0.382246188269289	0.452228401904335	0.414747227377831	0.523038749993159	0.392	-0.183781168551194	0.350964758875202	0.3932396767
-0.519172039299797	0.612485173196987	0.625680937548028	0.588350203621847	0.315213522924538	0.553204466635987	0.623489071902228	0.632436751849737	0.602603755530211	0.5861849374
-0.409941774604172	0.454031708986015	0.451046384304182	0.451780312199428	0.285343534484118	0.356669665906318	0.397443942456287	0.439985882991297	0.515055561088152	0.5585838599
-0.391226601821777	-0.028625802168888 9	-0.058539968233797 6	0.009537902789350 81	0.039300637979030 4	-0.043293829861367 6	-0.102284668134805	-0.075969794864514 2	0.026281754381271 5	0.1187731666
0.185888796030813	0.103858671338567	0.095765353873521 2	0.113356361516415	-0.030269158139258 7	0.125076952265285	0.078679805596753 7	0.108883011147021	0.122418159530867	0.1203010200
-0.345973651361771	0.574802373580257	0.569755137903114	0.573062324723844	0.238399723674502	0.496559534475463	0.557484791933011	0.562330326990801	0.611491231090022	0.6208770671
-0.263106753814212	0.259899477294985	0.234309586858847	0.287445391202078	0.328940175719376	0.232480716039598	0.186939041994 33	0.226924777541098	0.295332268029807	0.3198062031
-0.304610779640891	0.046685260748571 2	-0.016134605780020 8	0.128601833024 52	0.191508745970512	0.239133618986811	0.004068104995517 55	-0.054432965856934	-0.017177900720697 4	0.0778386458
-0.041963153714526 6	0.156295084474357	0.167256019506 68	0.139860741074 43	-0.006415720804478 23	0.051852068241078 5	0.163262850299 69	0.208534922633623	0.172916953923004	0.1652472873
-0.179445875487069	0.029212743434443 5	0.030609157188508 9	0.026883498272524 5	0.020967851157577 6	-0.017213744845225 2	-0.024503310370319 1	-0.000378573689628393	0.085820656376448 4	0.1697261201
0.130658435597814	-0.082695015785218 6	-0.046886385734067 4	-0.125696542530552	-0.219234476033321	-0.176800529194691	-0.044980803535156 5	-0.011308053934697	-0.064632707515285 1	-0.1072919721

<그림 7> 25개 구 & 45개 업종 상관관계 수치

2) 주성분분석, 변수제거법을 이용한 다중회귀분석

다중회귀분석은 업종별, 지역구별로 나누어 종속변수로는 월매출액을 선택하였고 독립변수로는 아파트(면적·가격별 세대 수), 유동인구(나이, 시간대, 성별, 요일), 주거인구(나이, 성별), 직장인구(나이, 성별), 매출(요일, 성별, 나이), 수입 및 지출(식품, 교육 등) 등 113개의 변수를 선택하였다.*

회귀의 기본가정 중 다중공선성의 문제를 해결하기 위해 두 가지 방법을 선택하였다.

첫 번째로 변수제거법을 선택하였다. R studio 기본패키지의 corr.test()함수와 'psych' 패키지의 pairs.panels 함수를 이용하여 직접 상관관계를 파악, 높은 변수 중 상관성이 높거나 의미를 가지는 변수를 직접 선택하는 방법을 선택하였다.

다른 방법으로는 step함수를 이용한 단계별 변수 선택 방법을 사용하였다.

두 번째로 주성분 분석을 prcomp함수를 이용하여 다중공선성의 문제를 해결하고 변수의 차원을 축소하였다. 주성분은 전체 데이터의 80% 이상을 설명할 수 있을 만큼의 개수를 선택하여 회귀분석을 하였다.

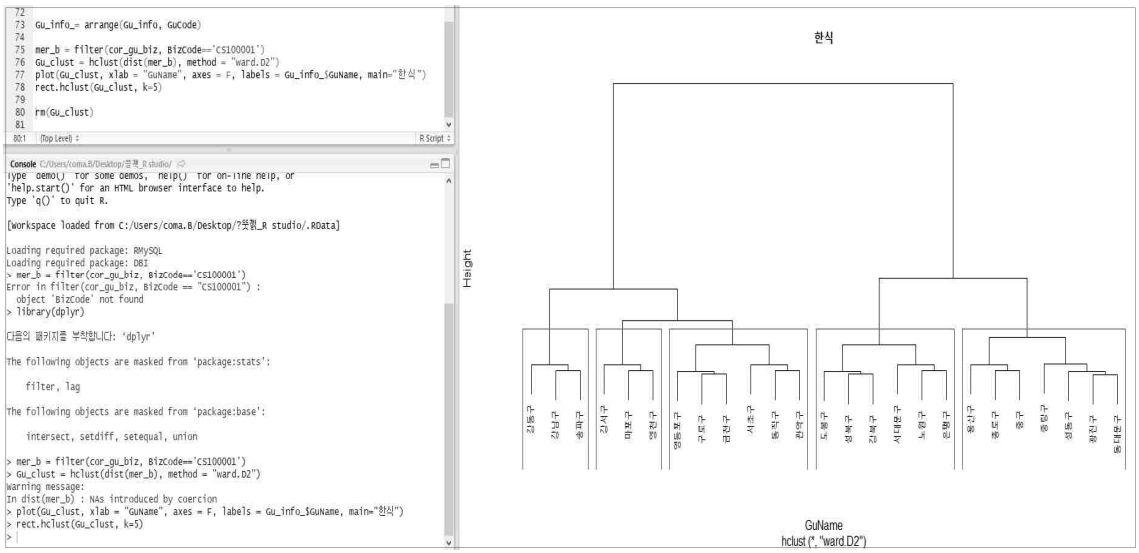
```
# 업종별 회귀식 CS100001
mer_a = filter(mer_v, BizCode=='CS100001')

mer_aa = mer_a[,-c(1:6,26,27)] # CodeNum, Date, CodeName, BizCode, BizName, Gucode, MonAveIncome, IncomeLevel

summary(mer_a)
boxplot(scale(mer_aa))
mer_a %>% arrange(desc(X20WorkPop))

#
lm.m=lm(SalesPerMon~., mer_aa)
summary(lm.m)
par(mfrow=c(2,2))
plot(lm.m)


#변수 추리기 step

step(lm(SalesPerMon~1, mer_aa), scope=list(lower=~1,
                                           upper=~APT_66Num+APT66Num+APT99Num+APT132Num+APT165Num+APTP_1Num+APTP1Num
                                           +APTP2Num+APTP3Num+APTP4Num+APTP5Num+APTP6_Num+X10FPop+X20FPop+X30FPop+X40FPop
                                           +X50FPop+X60FPop+X10Pop+X20Pop+X30Pop+X40Pop+X50Pop+X60Pop+AptNum+NAptNum+X10WorkPop
                                           +X20WorkPop+X30WorkPop+X40WorkPop+X50WorkPop+X60WorkPop), direction="both")
# CS100001 PCA
pcomp = prcomp(mer_aa, scale=T)
summary(pcomp)
pcomp$rotation
head(round(pcomp$x,3))
x = as.data.frame(round(pcomp$x,3))
idx = which(x$PC1==min(x$PC1))
mer_a[idx,]
biplot(pcomp)
```

<그림 8> 주성분분석, 변수제거법을 이용한 다중회귀분석 코드

* 선행논문(이명호, 2016)의 한계점을 극복하기 위해 다양한 변수를 선택하였다.

3) 클러스터 분석

클러스터 분석은 개체간의 유사성(가까운 거리)을 바탕으로 비슷한 특징을 나타내는 개체를 그룹화 하는 통계분석방법이다(Ceccato and Persson, 2002). 거리 행렬을 가지고 군집 간 유사성을 측정해서 모든 데이터가 하나의 군집으로 응집/병합될 때까지 반복적으로 군집화를 수행한다. 여러 거리측정방법 중 이상치에 덜 민감하고 비슷한 크기의 군집을 묶어주는 특징을 가진 와드(ward) 연결법을 사용하여 군집화를 진행하였다. <그림5>와 같이 군집이 형성되었고 상관분석을 통해 군집의 뚜렷한 특징을 파악하고자 군집간 거리를 고려하여 5개의 군집으로 나누었다.



<그림 9> 클러스터 분석 코드 및 군집

Ⅳ. 분석내용

1. 분석개요

서울시 골목상권과 상권배후지에서 공통적으로 분석한 대상은 지역구와 업종별 월매출액이다. 이들에 상관이 있는 요인을 분석하기 위해 선정한 변수는 크게 아파트, 유동인구, 세부매출, 집객시설, 상주인구, 소득소비, 점포, 직장인구 여덟 종류이다. 아파트에 해당하는 분석값은 단지 수, 평수, 가격, 평균면적과 시가이다. 이들 변수를 월매출액과 상관분석 한 결과 평수와 가격이 부분적으로 골목상권 및 배후지 월매출액과 상관관계가 있는 것으로 확인되었

다. 두 번째는 유동인구 데이터이다. 총 유동인구, 성별, 연령별, 시간대별, 요일별 유동인구를 매출과 상관분석 하였다. 이 결과 골목상권과 시간대별 유동인구만 제외하고** 대부분의 변수가 매출과 상관성이 있는 것으로 분석되었다. 세 번째 변수는 세부매출로 평균영업 개월, 생존율, 주중/주말 매출. 요일별, 성별, 연령별 매출이다. 평균영업 개월 수와 생존율만 제외한 대부분의 변수가 매출과 상관이 있는 것으로 분석되었다. 네 번째 집객시설은 골목상권과 배후지 모두 매출과 상관성이 적은 것으로 확인되었다. 상주인구는 총인구, 성별인구, 연령대별, 가구 수인데 대부분 매출과 상관이 있었다. 단, 골목상권에 대한 가구 수는 매출과 상관성이 적었다. 소득소비 규모는 골목상권과 배후지 모두 부분적으로 상관이 있는 것으로 확인되었다. 점포 수는 골목상권과 배후지 매출에 모두 관련이 없었다.*** 마지막으로, 직장인구는 성별, 연령별 변수 모두 골목상권 및 배후지 매출과 상관이 있는 것으로 확인되었다.

다중회귀분석을 위해 변수들 사이의 다중공선성을 먼저 확인한 결과, 유동인구, 세부매출, 상주인구, 소득소비, 직장인구의 변수들은 연령별, 시간대별, 요일별 등 세부적인 변수들 자체적으로 높은 상관성을 보였다. 따라서 총인구, 총금액 등 통합변수를 선택하거나 변수를 직접 제거하는 방법을 선택하였다. 하지만 회귀모델이 다중공선성의 문제로 인한 매우 높은 설명력(R^2 또는 $adjusted R^2$)을 갖거나 매우 낮은 설명력을 갖는 문제가 생겨 유의미한 결과를 도출하지 못하였다. 다른 방법으로 주성분분석을 이용하여 전체 데이터의 80%이상을 설명할 수 있는 주성분 4~5개를 선택하여 회귀모델에 적용하였다. 하지만 이 역시 다중공선성을 해결하지 못하고 98% 이상의 설명력(R^2값이 0.98 이상)을 갖는 문제가 있었다.**** 변수를 조정하며 주성분분석을 하였을 때도 약 20%의 설명력을 갖거나 극단적으로 98% 이상의 설명력을 갖는 결과를 얻어 유의미한 결과를 얻지 못하였다.*****

따라서 회귀분석을 제외한 상관분석과 군집분석을 통해 상권 분석을 진행하였다.

** 시간대별 유동인구는 지역구별 특징이 나타났지만, 군집화 한 지역구에서는 차이가 상쇄되는 경우가 있어 연구의 대상에서 제외하였다. 이 부분을 지역구별로 비교할만한 가치는 충분히 있다.

*** 매출테이블에 있는 점포 수 데이터와 점포테이블에 있는 점포 수 데이터가 상이하여 분석의 대상에서 제외하였다.

**** 다중공선성의 문제가 있을 경우 회귀 모델의 R^2(모델의 설명력)값이 매우 높게 나타나는 상황이 발생한다.

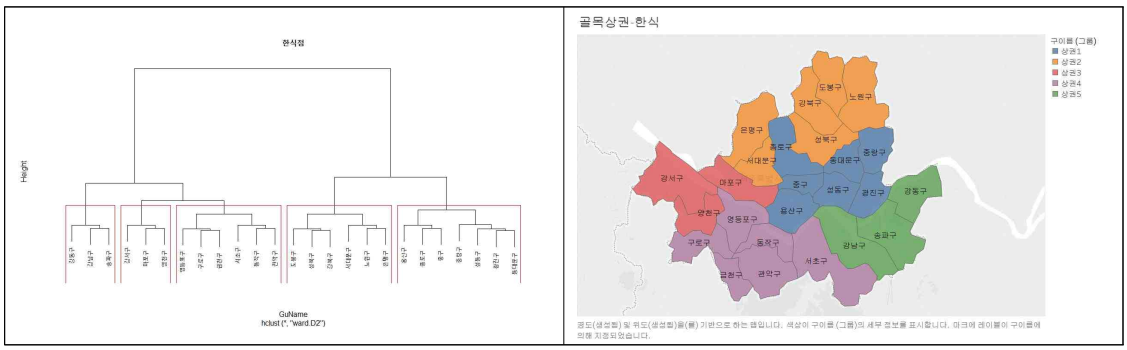
***** 분석에 소요되는 시간적 문제로 인하여 분석을 더 깊게 진행하지 못한 점은 본 연구의 한계점이다.

<표 1> 지역구별 업종별 월매출%의 분석개요

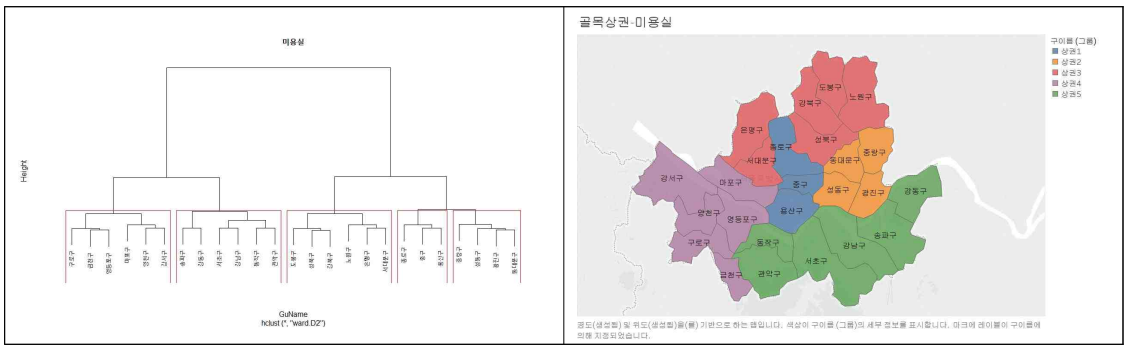
변수	분석값		단위	분석개요	
				골목상권	배후지
아파트	단지 수		단지	무의미	무의미
	평수	66미만, 66, 99, 132, 165 (m^2)	세대	부분적 유의미	부분적 유의미
	가격	1억 미만, 1억, 2억, 3억, 4억, 5억, 6억 이상	세대	무의미	부분적 유의미
	평균	면적	m^2	무의미	무의미
		시가	원	무의미	무의미
유동인구	총유동인구		명	유의미	유의미
	성별	남, 여	명	유의미	유의미
	연령	10, 20, 30, 40, 50, 60대 이상	명	유의미	유의미
	시간대	00~06, 06~11, 11~14, 14~17, 17~21, 21~24	명	무의미	부분적 유의미
	요일	월, 화, 수, 목, 금. 토	명	유의미	유의미
세부매출	평균영업개월 수		개월	무의미	무의미
	생존율	1년 이하, 1~2년, 2~3년, 3~5년, 5년 이상	%	무의미	무의미
	주중매출금액, 주말매출금액		원	유의미	유의미
	비율	주중, 주말, 월, 화, 수, 목, 금, 토, 일, 남, 여, 10, 20, 30, 40, 50, 60대 이상	%	무의미	무의미
	요일	월, 화, 수, 목, 금, 토, 일	원	유의미	유의미
	성별	남, 여	원	유의미	유의미
	연령	10, 20, 30, 40, 50, 60대	원	유의미	유의미
집객시설	시설(병원,관공서,은행 등) 개수		개	무의미	무의미
상주인구	총 상주인구		명	유의미	유의미
	성별	남, 여	명	유의미	유의미
	연령	10, 20, 30, 40, 50, 60대 이상	명	유의미	유의미
	성별,연령별	남:10, 20, 30, 40, 50, 60대 이상 여:10, 20, 30, 40, 50, 60대 이상	명	유의미	유의미
	총 가구		가구	무의미	부분적 유의미
	가구 수	아파트, 비아파트	가수	무의미	부분적 유의미
소득소비	월평균소득, 소득구간코드, 지출(총금액, 식료품, 의료 등) 총지출		원	부분적 유의미	부분적 유의미
점포	점포수	점포 수, 유사업종, 개업점포	개	무의미	무의미
	개업률, 폐업률		%	무의미	무의미
직장인구	총 직장인구 수		명	유의미	유의미
	성 별 , 연령별	남:10, 20, 30, 40, 50, 60대 이상 여:10, 20, 30, 40. 50, 60대 이상	명	유의미	유의미
	성별	남, 여	명	유의미	유의미
	연령	10, 20, 30, 40, 50, 60대 이상	명	유의미	유의미

2. 서울시 골목상권 · 상권배후지 군집분석

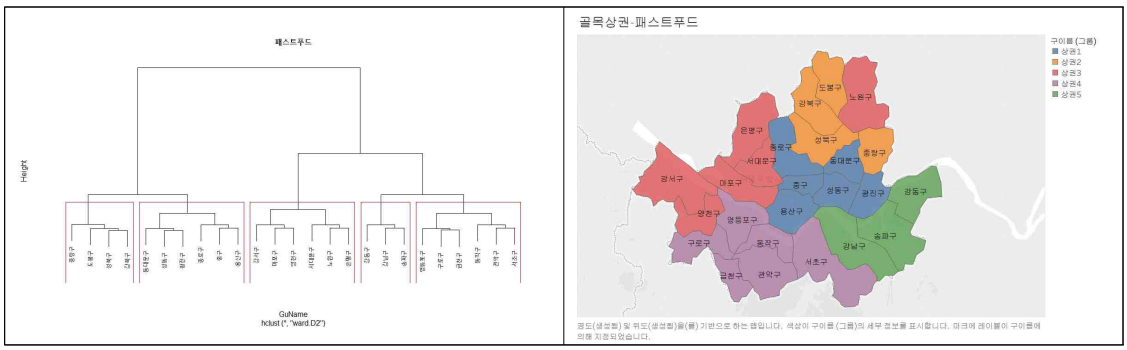
서울시 골목상권과 상권배후지에서 업종별 군집분석을 시행하였다. 그 결과 군집이 유사하게 형성되는 업종이 있지만, 차이가 있는 업종이 대부분이었다. 동일 업종이라도 골목상권과 상권배후지에서 다른 군집을 형성하였고, 같은 상권이라도 업종에 따라 군집이 다르게 형성되었음을 확인하였다. <그림6>~<그림13>은 한식, 미용실, 패스트푸드, 편의점의 상권별 군집이다.



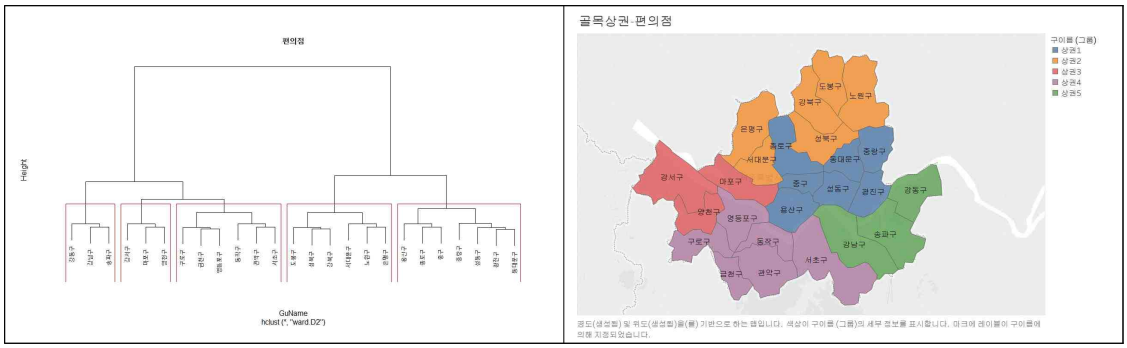
<그림 10> 서울시 골목상권 한식 군집분석



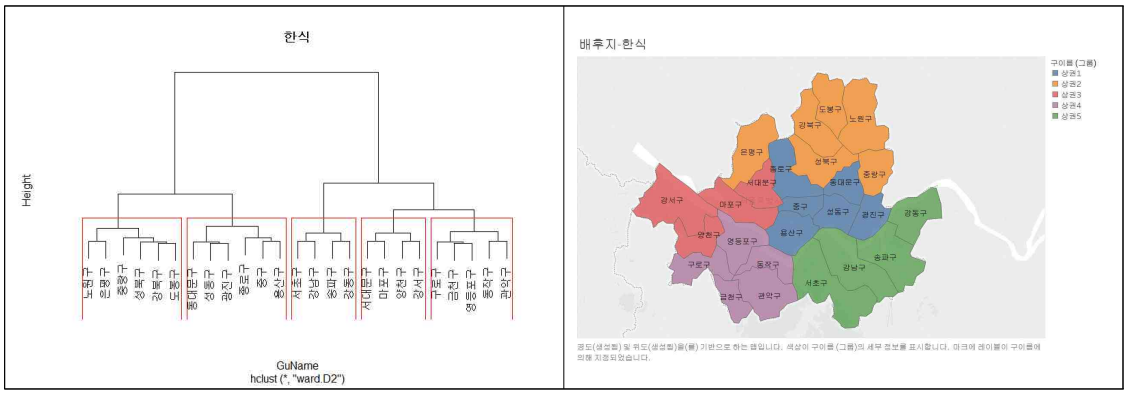
<그림 11> 서울시 골목상권 미용실 군집분석



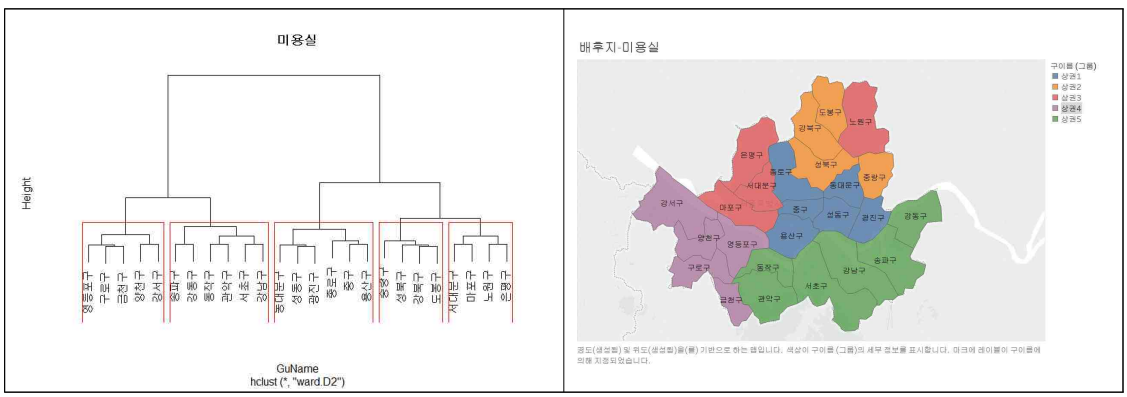
<그림 12> 서울시 골목상권 패스트푸드 군집분석



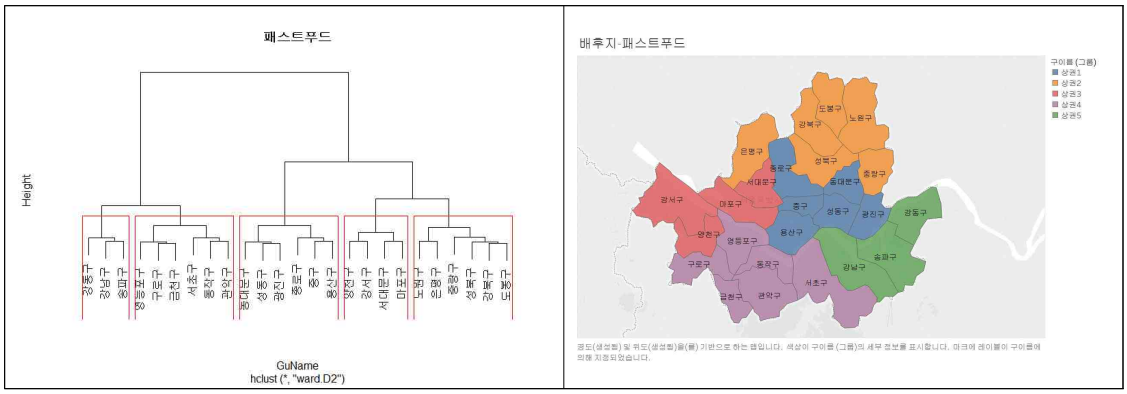
<그림 13> 서울시 골목상권 편의점 군집분석



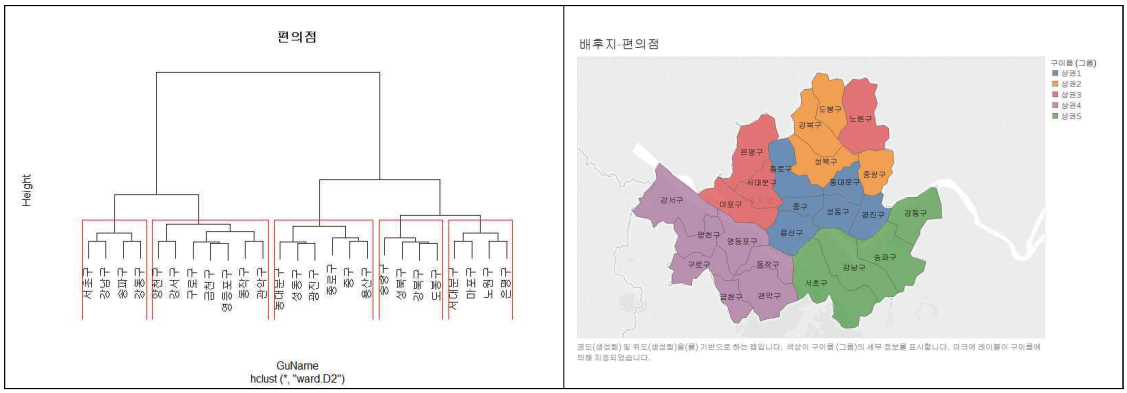
<그림 14> 서울시 상권배후지 한식 군집분석



<그림 15> 서울시 상권배후지 미용실 군집분석



<그림 16> 서울시 상권배후지 패스트푸드 군집분석



<그림 17> 서울시 상권배후지 편의점 군집분석

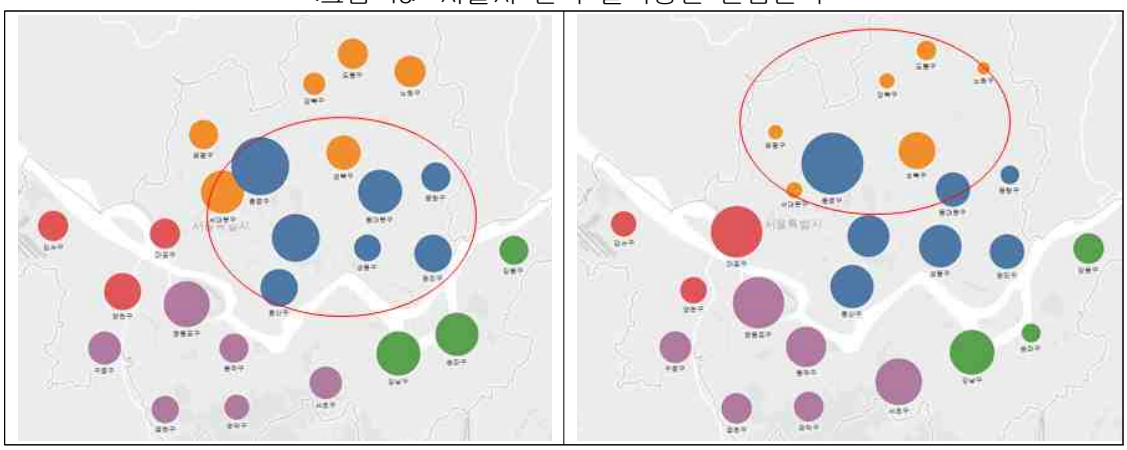
3. 유사 지역구들의 특징분석

1) 서울시 한식 골목상권 분석내용

한식의 골목상권 군집을 5개로 나누었을 때, 상권1은 종로구, 중구, 용산구, 성동구, 동대문구, 중랑구, 광진구로 구성되었다. 상권2는 성북구, 도봉구, 노원구, 은평구, 서대문구로 구성되었다. 상권3은 마포구, 강서구, 양천구로 구성되었다. 상권4는 구로구, 영등포구, 금천구, 동작구, 관악구, 서초구로 구성되었다. 상권5는 강남구, 송파구, 강동구로 구성되었다.

한식은 전반적으로 30~40대에 해당하는 변수들과 높은 상관관계를 가진다.

<그림 18> 서울시 한식 골목상권 군집분석



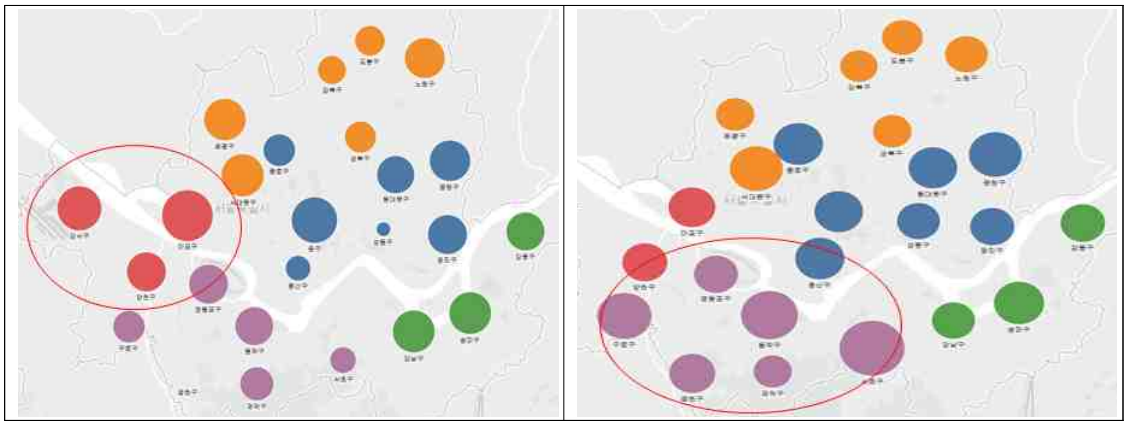
<그림 19 > 10~20대 직장인구(좌), 주중-주말 매출차이(우)

상권1(종로구, 중구, 용산구, 성동구, 동대문구, 중랑구, 광진구)은 '다른 연령에 비해 10~20대 직장인구와 높은 상관관계를 보였다. 그리고 주중 주말의 매출 상관계수의 차가 다른 지역에 비해 가장 컸다. 또한 '60' 대 거주인구 특징도 가지고 있다.

중구에서는 APT66m^2(0.47)이 다른 평수의 아파트보다 4배 이상의 상관관계를 보여주었다. 용산구(-0.6)와 성동구(-0.22)는 거주인구와 APT132m^2 이하에 '음의 상관관계' 를 보여주었고 132m^2이상에서는 '양의 상관관계' 를 나타냈다.

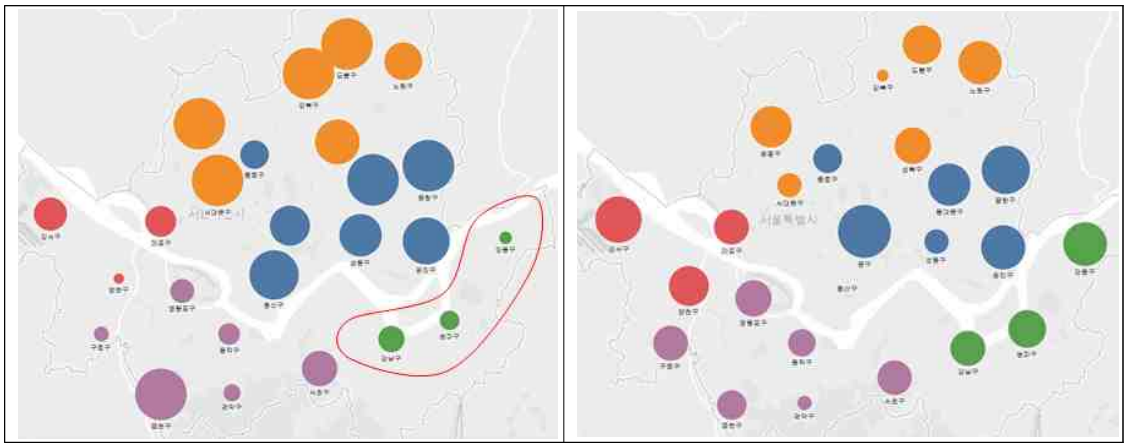
위의 결과로부터 추측해 보면 작은 평수에 사는 고령 인구의 특징을 내포하고 있는 지역이라는 통찰을 찾을 수 있었다.

상권2(성북구, 도봉구, 노원구, 은평구, 서대문구)는 편차가 나머지 지역에 비해 낮았고 10대를 제외한 전 연령의 특성이 두드러지게 나타났다. 유동인구에서만 '음의 상관관계'를 나타냈으며 거주 및 직장인구 에서는 비슷한 상관 수치를 나타냈다. 또한, 강북구는 거주인구에서 전 연령에 '음의 상관관계'를 나타냈다. 성북구는 APT66m^2 이상에서는 '음의 상관관계'를 보여줬다. 앞에서 나타난 편차와 지역적 특성을 합쳐보면 가족중심의 거주인구 특징과 지역 내에서 소비의 대부분이 발생하는 통찰을 확인 할 수 있었다.



<그림 20> 10대 거주인구(좌), 유동인구(우)

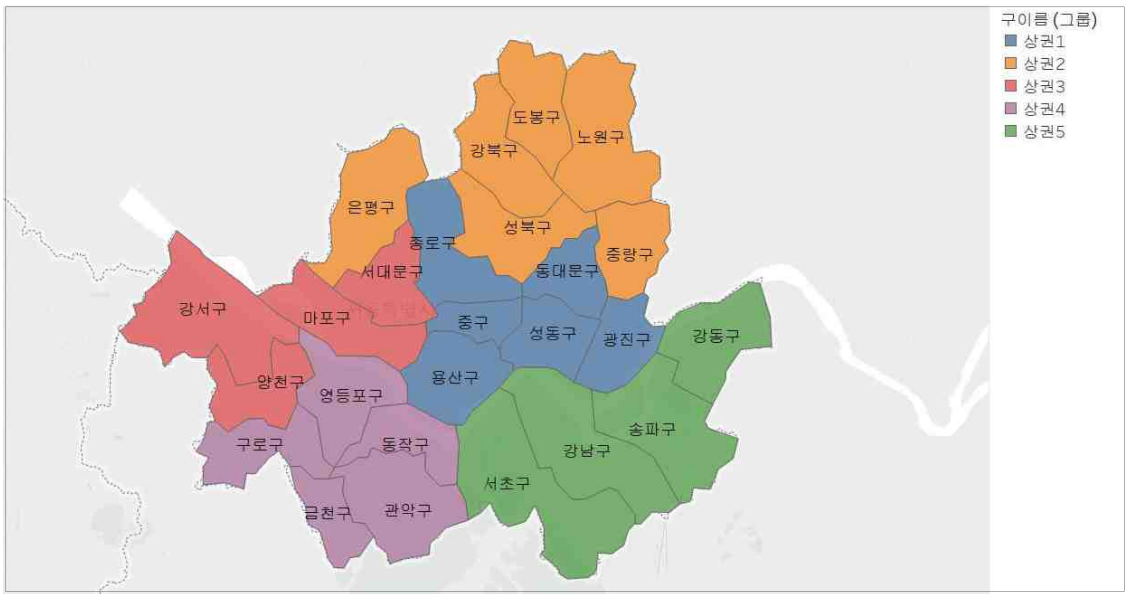
상권3(마포구, 강서구, 양천구)은 '10대' 특징을 나타냈다. 유동인구에서만 '음의 상관관계'를 나타냈고 거주인구 및 직장인구 에서는 양의 상관관계를 보여주었다. 특히 마포구 거주인구에서 '10대' 상관관계는 0.63으로 매우 높은 관계 값을 보여주었다. 양천구에서는 APT99m^2(0.35)와 강서구에서는 APT66m^2(0.35)와 상대적으로 높은 상관관계를 보여줬다. 상관계수를 통한 군집은 같은 지역으로 묶였지만, 마포구는 10대를 중심으로 하는 특성이 두드러지게 나타나는 통찰이 있었다. 상권4(구로구, 영등포구, 금천구, 동작구, 관악구, 서초구)는 20~40대의 특징을 나타냈다. 2~40대의 상관관계가 모든 지표에서 서로 비슷한 값을 나타냈다. 금천구는 거주인구의 전 연령에서 '음의 상관관계'(-0.33)로 높은 수치를 나타냈다. 구로구, 금천구, 영등포구, 동작구는 APT165m^2와는 '음의 상관관계'를 보이는 특성을 나타냈다. 20~40대 소비중심과 지역적 특성을 통해 상권4는 개인 혹은 2인 가족의 특성을 보유하고 이를 통한 상권이 존재하는 통찰을 확인 할 수 있었다.



<그림 21> 고가 아파트(좌), 저가 아파트(우)

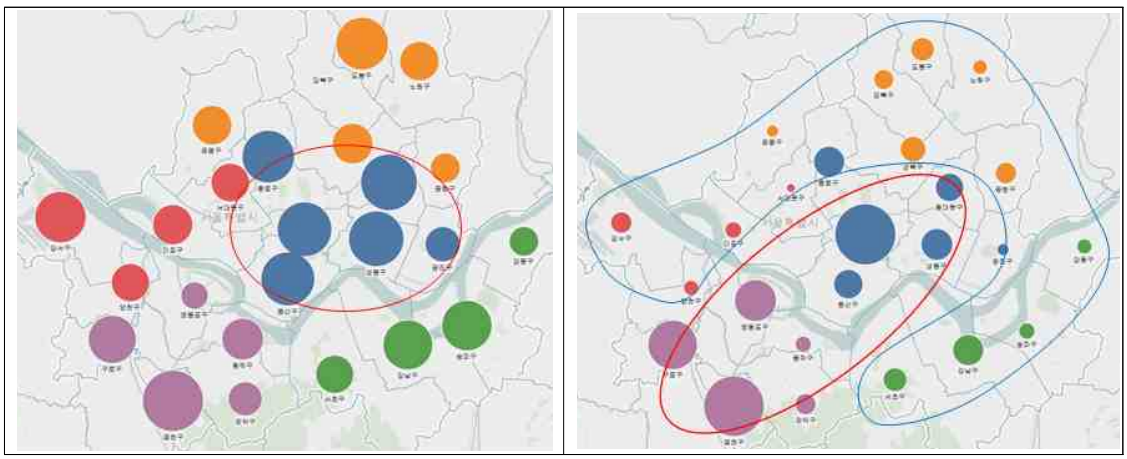
상권5(강남구, 송파구, 강동구)는 고가 아파트와는 상관관계가 낮은 것으로 나타나며 전 연령의 특징이 비슷한 상관관계를 보여주며 특히 거주인구와 직장인구가 상대적으로 높은 '양의 상관관계'를 보여준다. 또한 4~50대 매출이 발생하면서도 10대의 인구 특징이 나타난 특성을 나타냈다. 이를 통해서 10대를 부양하는 4~50대 가족이 구성되어 상권이 형성되고 있다는 통찰을 해볼 수 있었다.

2) 서울시 한식 상권배후지 분석내용



<그림 22> 서울시 한식 상권배후지 군집분석

한식의 상권배후지를 5개 군집으로 나누었을 때, 상권1은 종로구, 중구, 용산구, 성동구, 광진구, 동대문구로 구성되었다. 상권2는 중랑구, 성북구, 강북구, 도봉구, 노원구, 은평구로 구성되었다. 상권3은 서대문구, 마포구, 강서구, 양천구로 구성되었다. 상권4는 구로구, 금천구, 영등포구, 동작구, 관악구로 구성되었다. 상권5는 서초구, 강남구, 송파구, 강동구로 구성되었다.

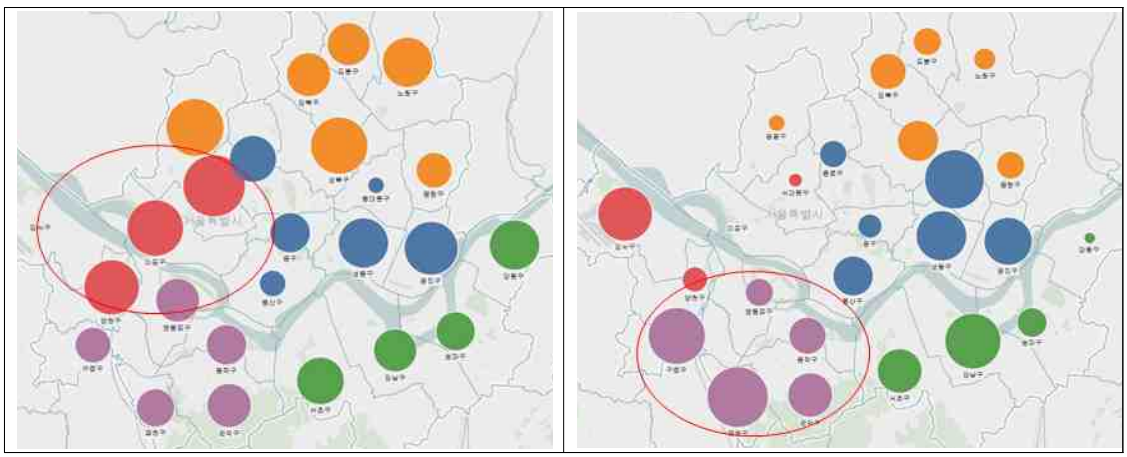


<그림 23> 총 직장인(좌), 주중-주말 매출차이(우)

상권1(종로구, 중구, 용산구, 성동구, 광진구, 동대문구)은 직장인과 상관관계가 높은 지역이다. 이 지역에서의 매출액은 다른 상권에 비해 직장인 인구와 양의 상관관계(0.41)가 높게 나타났다. 특히, 30대(0.41), 40대(0.4), 50대(0.37) 직장인구의 상관계수가 나머지 4개 상권의 평균과 약 2배 차이를 보였다.

또한, 이 지역은 매출과 유동인구의 상관관계가 다른 지역에 비해 낮은데 이는 매출에 기여하지 않는 유동인구가 많은 지역임을 추측할 수 있다. 반면, 직장인인구와는 다른 지역에 비해 가장 높은 상관성을 보였다(0.41로 다른 지역의 약 2배).

상권2(중랑구, 성북구, 강북구, 도봉구, 노원구, 은평구)는 변수 간 균형이 잡혀있는 지역이다. 이 지역에서의 매출액은 다른 지역 상권보다 주중-주말, 연령별 매출액, 유동인구, 직장인구 상관계수를 비교하였을 때 가장 낮은 편차를 가진다. 즉, 특정 요일이나 특정 연령에 관계없이 고르게 매출이 발생한다고 설명할 수 있다. 또한, 이 지역에서는 다른 연령에 비해 60대 유동인구와의 상관계수가 가장 높은 양의 상관관계를 보였다(60대>20대>50대). 이는 다른 5개 상권의 20, 30, 40대 유동인구 상관계수가 높다는 점과 차이가 있다.



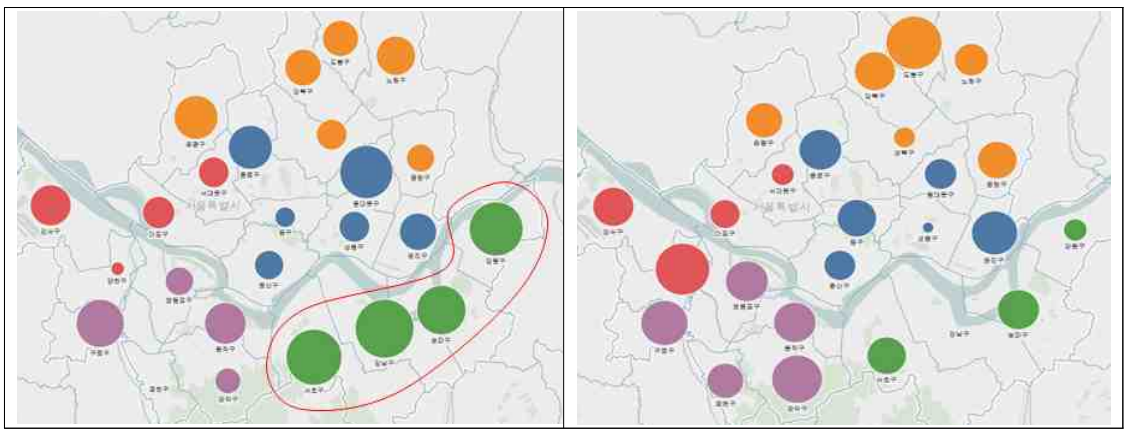
<그림 24> 10~20대 유동인구(좌), 남-녀 유동인구(우)

상권3(서대문구, 마포구, 강서구, 양천구)은 유동인구, 특히 젊은 세대와 상관관계가 높은 지역이다. 이 지역에서의 매출액은 다른 지역에 비해 10대(0.32), 20대(0.53) 유동인구와 상관계수가 높은 편이다. 그리고 거주인구와는 음의 상관관계를 가진다. 특히, 서대문구(-0.44)와 마포구(-0.54)의 거주인구는 음의 상관관계를 가진다. 이를 통해, 이 지역에서의 매출은 거주인구보다는 유동인구와 양의 상관성이 높다고 설명할 수 있다.

특히 양천구에서는 아파트 상관계수 차이가 극명하게 드러나는데, 아파트 66m^2미만(-0.54), 132m^2(0.51), 165m^2(0.52)와는 강한 양의 상관관계가 있다. 그리고 2억 이하 아파트(약 -0.3)와는 음의 상관관계인 반면 3억 이상 아파트(약 0.4)와는 양의 상관관계가 있다. 이로써 상권3에서 좁은 면적 아파트보다는 넓은 면적 아파트와, 저가 아파트보다는 고가 아파트와 높은 양의 상관성을 띔을 알 수 있다.

상권4(구로구, 금천구, 영등포구, 동작구, 관악구)는 남성, 직장인과 높은 상관성을 띄는 지역이다. 주중매출(0.98), 주말매출(0.80)이 다른 지역에 비해 차가 가장 심한 지역이다. 특히 금천구는 상관계수가 주중(0.96), 주말(0.56)이며, 구로구는 주중(0.98), 주말(0.72)로써 월매출이 주말보다는 주중과 상관성이 높음을 알 수 있다.

또한, 남녀 매출(남성 0.97, 여성 0.91) & 유동인구(남성 0.44, 여성 0.36) 차이가 다른 지역에 비해 가장 높았다. 이를 통해 여성보다는 남성과 상관성이 높음을 알 수 있었다. 또한 직장인구와는 0.28의 양의 상관관계를 가진다. 특히, 30대와 40대에서는 남성이 여성보다 높은 상관계수를 가짐을 확인할 수 있었다. 이를 종합해 보았을 때, 주중, 남성, 30대, 40대의 특징을 뽑아내었고, 이 지역에 IT회사가 밀집해 있다는 점을 고려해 보았을 때 특징이 이 지역의 대표성을 설명할 수 있다고 판단하였다.

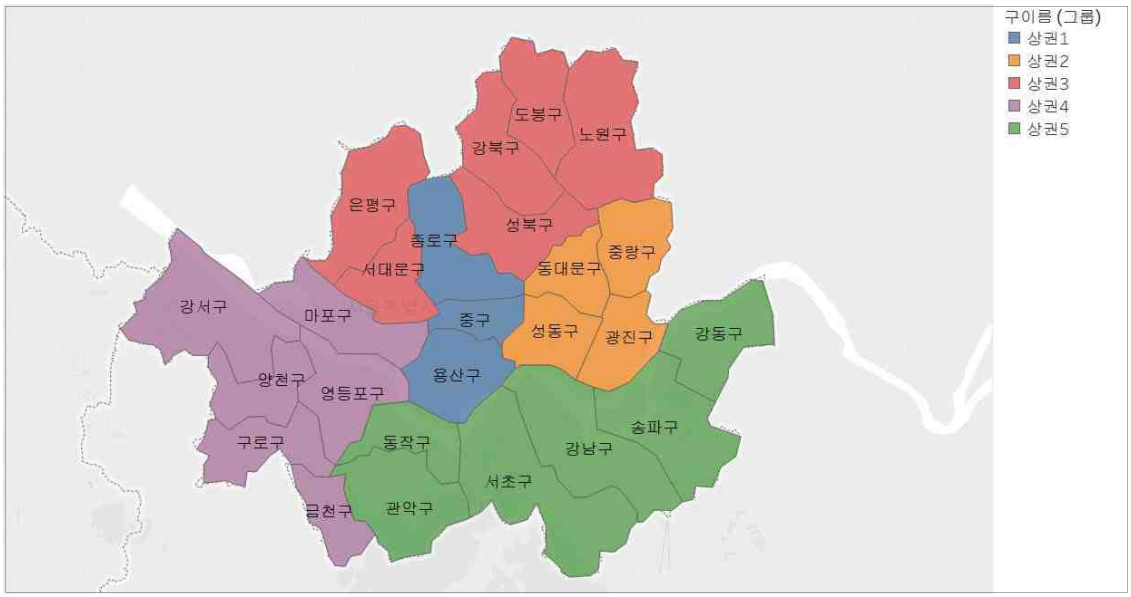


<그림 25> 저가 아파트(좌), 고가 아파트(우)

상권5(서초구, 강남구, 송파구, 강동구)는 부동산과 상관관계가 높은 지역이다. 이 지역은 다른 지역에 비해 세대 수, 가구 수와 관련이 높았다. 우선, 아파트 단지 수(0.23)와 양의 상관관계를 가지지만 다른 지역은 음의 상관관계를 보여 대비되는 성향을 보였다(예외적으로 동대문구(0.44), 구로구(0.34), 동작구(0.24)의 양의 상관관계). 또한, 비아파트(0.3)는 양의 상관관계를 가졌는데 다른 지역의 비 상관성과 비교해 상권5만의 특징이다.
그리고 $66m^2$미만 아파트(0.23)는 양의 상관관계를 가진 반면 그 이상의 아파트와는 음의 상관이거나 거의 없었다. 가격별로는 1억 미만(0.41), 1억(0.29)이 양의 상관을 띄었고 이상의 아파트에 대해서는 거의 없거나 음의 상관관계를 띄었다.
유동인구와의 상관관계는 다른 지역에 비해 가장 강한 양의 상관관계를 띄지만(0.49), 특이하게 10대 유동인구와는 0.06의 낮은 상관계수를 갖는 것이 이 지역의 특징이다.
거주인구와(0.27) 양의 상관을 보이는데, 다른 지역이 음의 상관 또는 상관관계가 거의 없는 것을 감안할 때, 이 지역만의 명확한 특징이라고 볼 수 있다. 특히, 다른 연령대에 비해 20대 거주인구(0.34), 30대 거주인구(0.39)와 강한 양의 상관을 갖는다.

3) 서울시 미용실 골목상권 분석내용

미용실의 골목상권을 5개 군집으로 나누었을 때, 상권1은 종로구, 중구, 용산구로 구성되었다. 상권2는 성동구, 광진구, 동대문구, 중랑구로 구성되었다. 상권3은 성북구, 노원구, 도봉구, 강북구, 은평구, 서대문구로 구성되었다. 상권4는 마포구, 강서구, 양천구, 구로구, 금천구, 영등포구로 구성되었다. 상권5는 동작구, 관악구, 서초구, 강남구, 송파구, 강동구로 구성되었다. 미용실을 전 변수들에서 평균적으로 '양의 상관관계'를 나타냈다. 특히 10~40대까지의 지표들이 전 지역에서 높은 상관관계를 나타냈다. 그리고 모든 지표에서 남성보다는 여성과 높은 상관관계를 보여줬다.



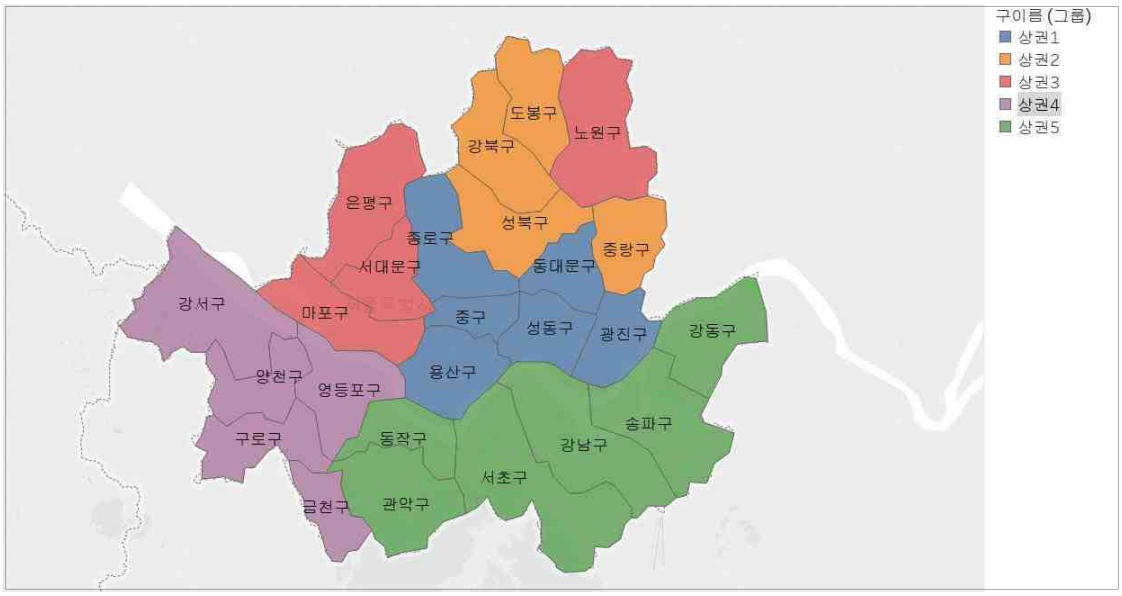
<그림 26> 서울시 미용실 골목상권 군집분석

반대로 종로구는 전 인구지표에서 '음의 상관관계'를 나타냈다. 금천구는 유동인구와 거주인구에서 '음의 상관관계'를 나타냈으며 종로구와 성동구는 직장인구 에서 -0.28과 -0.22로 낮은 상관관계를 나타냈다. 용산구는 유동인구에서 전체적으로 높은 상관관계(10대:0.67, 20대:0.62, 30대:0.35, 40대:0.37, 50대:0.35, 60대:0.33)를 보여줬으며 광진구는 거주인구에서 높은 상관관계(10대:0.42, 20대:0.62, 30대:0.54, 40대:0.47, 50대:0.45, 60대:0.49)를 나타냈다. 요일별 매출에서는 토요일이 가장 상관관계가 높았으며 그다음은 금요일로 2 개 구를 제외한 대부분 구에서 높은 상관관계를 보여준다.

4) 서울시 미용실 상권배후지 분석내용

미용실의 상권배후지를 5개 군집으로 나누었을 때, 상권1은 종로구, 중구, 용산구, 성동구, 광진구, 동대문구로 구성되었다. 상권2는 중랑구, 성북구, 강북구, 도봉구로 구성되었다. 상권3은 노원구, 은평구, 서대문구, 마포구로 구성되었다. 상권4는 강서구, 양천구, 구로구, 금천구, 영등포구로 구성되었다. 상권5는 동작구, 관악구, 서초구, 강남구, 송파구, 강동구로 구성되었다. 상권배후지에서 미용실업종의 매출은 전반적으로 직장인인구와는 상관성이 매우 낮지만 거주인구, 유동인구와는 양의 상관관계를 가진다. 상권1(종로구, 중구, 용산구, 성동구, 광진구, 동대문구)에서의 매출은 다른 연령대에 비해 10대~30대 유동인구와 양의 상관성이 높았다(20대>10대>30대>40대>60대>50대로써 각각 0.44, 0.37, 0.28, 0.24, 0.22, 0.21). 또한, 중구는 다른 지역에 비해 직장인 인구와 가장 높은 양의 상관성(0.38)을 보였다. 그중에서도 20~40대 직장인인구와 0.41, 0.44, 0.46의 상관계수로써 다른 연령대 평균인 0.21보다

높았다. 반면 종로구는 40~60대 직장인 인구와 각각 0.19, 0.28, 0.38의 상관계수로써 다른 연령대 평균인 -0.06보다 높은 양의 상관성을 띄었다. 또한, 종로구는 총가구 수, 세대 수와 높은 상관관계를 가진다(총 가구 수, 아파트 세대 수, 비아파트 세대 수 순으로 각각 0.76, 0.75, 0.65).



<그림 27> 서울시 미용실 상권배후지 군집분석

상권2(중랑구, 성북구, 강북구, 도봉구)에서의 매출은 다른 연령대에 비해 50~60대 유동인구와 양의 상관관계가 높았다(60대>50대>40대>20대>10대>30대로써 각각 0.46, 0.43, 0.41, 0.41, 0.40, 0.40). 또한 이 지역은 다른 요일에 비해 수요일과 목요일 유동인구와 상관관계가 높다. 이는 다른 4개의 상권매출이 토요일과 일요일 유동인구와 상관관계가 높은 것과는 다른 결과이다.

상권3(노원구, 은평구, 서대문구, 마포구)은 상권1과 같이 10~30대 유동인구와 양의 상관관계가 높은 지역이다(20대>10대>30대>40대>50대>60대로써 각각 0.68, 0.60, 0.54, 0.48, 0.48, 0.45). 또한, 이 지역은 주말과 주중 간 매출 상관계수 차가 0.005로써 5개 상권 중 가장 작다. 그리고 연령별 매출액 상관계수의 표준편차가 0.05로써 종합하여 볼 때, 전 연령에 걸쳐 주중 주말과 관계없이 골고루 매출이 발생함을 알 수 있다.

다른 4개의 상권이 총 가구 수, 아파트, 비아파트 세대 수, 거주인구와 양의 상관관계를 띄지만, 이 지역에서는 유일하게 음의 상관관계를 나타내었다. 대신, 이 지역은 유동인구와 가장 높은 양의 상관관계(0.60)를 나타낸다. 종합하여 볼 때, 상권3에서의 미용실 매출은 거주인구보다는 유동인구와 높은 양의 상관관계를 나타낸다고 할 수 있다. 특히, 마포구와 서대문구는 거주인구와 각각 -0.42, -0.36의 비교적 명확한 음의 상관성을 보였다.

상권4(강서구, 양천구, 구로구, 금천구, 영등포구)는 주중과 주말 매출 상관계수의 차가 다

른 지역에 비해 심한 지역이다. 특히, 구로구, 금천구, 영등포구는 다른 업종(한식, 편의점 등)에서와 같이 주중과 주말의 매출 상관계수의 차가 심하게 나타났다. 또한, 10대 거주인구와의 상관성이 다른 지역에 비해 높았다(상권4>상권5>상권1>상권2>상권3 순으로 각각 0.40, 0.21, 0.21, 0.18, -0.24). 이 지역에서는 저가보단 고가 아파트 세대 수와 양의 상관관계를 가지고 마찬가지로 거주인구와도 양의 상관관계를 가진다. 특히, 영등포구, 양천구에서는 1억 이하 아파트 세대 수와 음의 상관관계가 있었지만, 3억 이상 아파트 세대 수와는 양의 상관관계가 있었다(영등포구의 경우, 1억 미만 -0.30, 1억 -0.31, 3억 0.54, 4억 0.55, 5억 0.50, 6억 이상 0.46). 금천구에서는 직장인구와 -0.21의 음의 상관관계가 있었다. 이는 한식과 0.59의 양의 상관관계를 갖는 것과는 상반된 금천구 미용실업종의 특징이다.

상권5(동작구, 관악구, 서초구, 강남구, 송파구, 강동구)는 비아파트 세대 수와 양의 상관관계(0.31)를 가지는데 동 지역 아파트 세대 수(0.11)보다 높고, 다른 상권에 비해서도 높은 양의 상관성을 띈다. 또한, 거주인구와 0.33의 양의 상관관계를 가진다.

Ⅴ. 결론

1. 분석결과

분석결과 연구 질문에 대한 대답은 다음과 같다. Q1.서울시 골목상권과 상권배후지의 업종별 지역구별 매출에 미치는 요인은 무엇인가?" → 서울시 골목상권과 상권배후지의 월매출과 상관성이 있는 변수는 아파트 단지 수(평별, 가격별), 매출(연령별, 요일별, 성별), 유동인구(연령별, 요일별), 연령별 거주인구, 직장인구(연령별, 성별)이다. 반면, 시간대별 유동인구, 소득 및 지출액, 성+연령별 거주인구, 성+연령별 직장인 인구는 큰 차이가 없거나 유의미하지 못하였다. Q2. 서울시 골목상권과 상권배후지의 업종별 매출에 미치는 요인이 유사한 지역구는 무엇인가? 골목상권과 상권배후지의 상권지역 군집의 특징은 대부분 서로 붙어있는 지역구끼리 붙어서 군집을 이루는 특징을 보여주고 있다. 이 중 동일하게 포함된 지역구를 표시했다.

Q3. 서울시 골목상권과 상권배후지의 업종별 유사한 지역구들의 특징은 무엇인가?

업종이 한식인 골목상권에서 상권1 지역은 노인 거주자와 연관된 특징을 보여준다. 상권2 지역은 모든 변수에서 양의 상관관계를 가지며 대부분 지표에서 편차가 가장 작게 나타나는 특징이 있다. 상권3 지역은 10대의 거주인구 및 유동인구와 연관된 특징이 있다. 상권4 지역은 20~40대 연령의 거주인구 및 직장인구와 연관된 특징이 있다. 상권5 지역은 10대와 40~50대 인구특징이 나타나고 매출은 40~50대 관계가 높이 나타나는 특징을 보여줬다.

업종이 한식인 상권배후지에서 상권1 지역은 20~40대 거주 및 직장인구와 연관된 특징이 나타났다. 상권2 지역은 전 연령에 걸쳐 대부분 변수의 상관관계 편차가 작게 나타나는 특징이 있었다. 상권3 지역은 10~20대 유동인구와의 관계가 특징적으로 나타났다. 상권4 지역은 주중과 주말 매출 상관계수의 차가 가장 높았고, 남자가 여자보다 매출에 높은 관계 값을 보여주

는 특징을 가졌다. 상권5 지역은 20~30대 거주인구의 특징과 부동산 관련 변수와 높은 상관성이 있는 특징이 있었다.

<표2> 상권구성

업종	상권구성
한식	상권1 : 종로구, 중구, 용산구, 성동구, 동대문구, 광진구 상권2 : 성북구, 강북구, 도봉구, 노원구, 은평구 상권3 : 마포구, 강서구, 양천구 상권4 : 영등포구, 구로구, 금천구, 관악구, 동작구 상권5 : 강남구, 송파구, 강동구
미용실	상권1 : 종로구, 중구, 용산구 상권2 : 중랑구 상권3 : 노원구, 은평구, 서대문구 상권4 : 강서구, 양천구, 구로구, 금천구, 영등포구 상권5 : 동작구, 관악구, 서초구, 강남구, 송파구, 강동구

업종이 미용실인 골목상권에서 상권1 지역은 10대 거주인구와 유동인구가 특징으로 나타났다. 상권2 지역은 20대 연령이 주로 높은 관계를 가지는 특징을 나타냈다. 상권3 지역은 전 연령에 걸친 특징이 고루 나타나 특별한 관계가 없는 대신 편차가 적은 특징을 나타냈다. 상권4 지역은 10~20대 거주인구 및 유동인구와 관계된 특징을 나타냈다. 상권5 지역은 20~30대 인구와 연관된 관계를 보여주는 특징을 지녔다.

업종이 미용실인 상권배후지에서 상권1 지역은 10~30대 유동인구와 연관된 특징을 나타냈다. 상권2 지역은 50~60대 유동인구와 연관된 특징을 보여주며 특이하게도 수, 목요일과 연관되어 관계가 높게 나타났다. 상권3 지역은 10~20대 유동인구와 연관된 특징을 가졌다. 상권4 지역은 10대의 거주인구와 연관된 특징을 가지며 고가 아파트와도 높은 관계를 나타냈다. 상권5 지역은 20~40대 거주인구와 관계된 특징을 지니고 비아파트 세대와 관계된 특징이 높게 나타났다.

2. 함의점

한식과 미용실 업종별 지역구의 매출을 분석한 결과 큰 함의점을 발견하였다. 그것은 업종별로 서울시의 골목상권은 특정한 지역구가 유사한 특성을 보인다는 점이다.

먼저 한식의 골목상권과 배후지 상권분석결과를 보면 종로구 중심의 상권1은 노인 인구가 중요하지만 배후지 상권은 직장인구가 중요한 요인이었다. 성북구 중심의 상권2는 골목상권에서는 주거인구가 중요했고, 배후지에서는 다른 구와 비교해 연령 성별 모두 균형적인 매출 상관 요인을 가지고 있었다. 마포구 중심의 상권3은 골목상권과 배후지 모두 10대 또는 젊은 인구가 상권매출액과 상관이 있었다. 영등포 중심의 상권4는 직장인과 남성 인구가 골

목상권과 배후지 매출과 상관이 있었다. 강남구 중심의 상권5는 10대와 40대의 인구가 골목상권 매출액과 상관이 있고, 배후지 상권과 상관성이 높은 요인은 부동산 변수였다.

미용실 업종에 대한 골목상권 분석은 다음과 같다. 종로구 중심의 상권1은 골목상권과 배후지 모두 10대 인구와 상관이 있었다. 중구 중심의 상권 2는 골목상권은 20대 인구가 중요하나 배후지는 오히려 50, 60대 인구가 중요하였다. 노원구 중심의 상권3의 골목상권은 10~20대 유동인구, 20~30대 거주인구와 상관이 있었지만, 상권배후지에서는 유동인구와 높은 상관성이 있었다. 강서구 중심은 상권4는 10, 20대 인구가 골목상권과 높은 상관성이 있고 배후지 매출은 고가 아파트와 상관이 있었다. 강남구 중심의 상권5는 20, 30대가 골목상권과 상관성이 높았고, 비아파트 거주 가구 수가 배후지 매출과 상관성이 컸다.

<표3> 상권별 한식, 미용실 골목상권 분석 지표

업종	상권	상권1	상권2	상권3	상권4	상권5
한식	지역구	종로구, 중구, 용산구, 성동구, 동대문구, 광진구	성북구, 강북구, 도봉구, 노원구, 은평구	마포구, 강서구, 양천구	영등포구, 구로구, 금천구, 관악구, 동작구	강남구, 송파구, 강동구
	골목상권	10&20대 직장인, 노인, 거주자	주거지&균형	10대, 거주인구	직장인,유동인구	10대, 40대, 고가 아파트
	상권배후지	직장인	균형	젊음	남성	부동산
미용실	지역구	종로구, 중구, 용산구	중랑구	노원구, 은평구, 서대문구	강서구, 양천구, 구로구, 금천구, 영등포구	동작구, 관악구, 서초구, 강남구, 송파구, 강동구
	골목상권	10대 거주&유동인구	20대	10~30대	10대 20대 거주&유동인구	20대, 30대
	상권배후지	10~30대 유동인구, 직장인	50대, 60대 유동인구	유동인구	고가아파트	비아파트, 거주

이러한 분석결과로 볼 때 서울시 골목상권 활성화를 위해선 업종별 지역별로 다른 요인을 사업전략의 기준으로 삼아야 한다는 것을 확인할 수 있다. 서울시는 골목상권지표를 제공하고 있는데, 지역구별 업종별로 특화된 것이 아니라 지역구의 인구통계학적인 변수를 획일적인 방법으로 적용하고 있다. 이 연구에서 도출한 결과와 같이 전체 업종별 매출요인을 해마다 분석하여 서울시 소상공인에게 제공한다면 서울시 지역발전과 서울시의 공공정책 발전에도 도움이 되리라 기대한다.

3. 한계 및 향후과제

본 연구의 한계점은 첫째, 데이터의 수집 시기와 갱신주기가 달라 갱신주기가 긴 변수는 일정 기간 내 같다는 가정에 따라 연구를 진행했다는 것이다. 둘째, 대외비 데이터를 공개 데이터와 함께 분석모델에 적용할 수 없었다는 것이다. 셋째, 연구 기간이 촉박하여 여러 방면에서 접근하는 것이 제한된 것이다.

데이터의 한계 상 2015~2016년의 데이터만을 가지고 분석을 진행하였는데, 더 많은 데이터가 있다면 시계열 분석을 통해 상권의 변화과정도 관찰할 수 있을 것이다.

참고문헌

구자용, “공간정보 빅데이터의 지도화와 공간적 분포 특성에 관한 연구”, *국토지리학회지*, 제49권, 3호, 2015, pp.349-360.

이경주, 홍성조, 고석관, *공간가중회귀모형을 이용한 강원도 지역상권 현황 진단 및 발전방안 연구*, 한국은행 강릉본부, 2015.

이명호, *공간 빅데이터를 활용한 소지역 상권 매출에 영향을 미치는 요인분석에 관한연구*, 안양대학교 대학원, 2016.

이승주, “주성분 분석을 이용한 빅데이터 분석”, *한국지능시스템학회 논문지*, 제25권, 6호, 2015, pp. 592-599.

이재윤, “지적 구조 분석을 위한 새로운 클러스터링 기법에 관한 연구”, *한국정보관리학회*, 2006, pp. 215-231.

정대석, 김형보, “상권 업종별 분포 및 매출 영향요인 분석”, *GRI 연구논총*, 제16권, 2호, 2014, pp. 101-122.

최재혁, 신창섭, “빅데이터를 활용한 휴양림 이용객현황과 인터넷 검색어의 상관관계분석”, *한국산림휴양학회*, 2015, pp. 13-23

Ceccato, V. and L. O. Persson, “Dynamics of rural areas: an assessment of clusters of employment in Sweden”, *Journal of Rural Studies,* Vol.18, 2002, pp 49-63.

K. Pearson, "On lines and planes of closest fit to systems of points in space", *Phil Mag*, vol.2, 1901, pp.559-572.

Analysis of market district hinterland in Seoul by 'Big Data' analysis

HeungRok Oh*, BumSeok Bae**, ***HyeJung Moon

Abstract

The Seoul government has opened the commercial data and analysis results of the Seoul market districts. Most of the data are consists of statistics. The researchers hope to derive new analysis results from the data. Therefore, the subject of research is 1008 defined commercial districts and hinterland from 2015 to 2016. Topics are apartment buildings, housing, facilities, incomes, sales of industries, population (floating, resident, employee), etc.

The research procedure follows 1. Collecting the public data in Seoul, 2. Building the database system for analysis using MySQL, 3. Data analysis using MsExcel 4. Performing correlation & regression analysis by R studio, 5. Using Tableau tool for visualization.

The analysis results show that Sales per month is associated with apartments size, population(resident, floating, employee), daily sales, sales per ages and expenditures. 25 administrative districts of Seoul were devided into 5 sectors by clustering using correlation coefficients. The community consisted mainly of neighborhood districts, and the sale of community stores in Gangnam was highly related with the number of lower priced apartments in the apartment complex. On the other hand, Dobong-gu is kind of balanced market.

Hence, we could identify the components of the Seoul metropolitan county, which is formed differently according to the type of industry, and identify characteristics and sales correlation factors of each community.

Key words: Big Data, Correlation Analysis, Cluster Analysis, Market district, Hinterland

* heungrokoh@gmail.com Bachlor of Mathematics, Thejoeun IT Academy
** qoqjatjr10@naver.com Bachlor of Computer Communication Information, Thejoeun IT Academy
*** Corresponding Author, hyejung.moon@gmail.com Ph D. of Public Policy, Adjunct Professor in Seoul University of Science & Technology, President of ILP, Director of Will-Be Solution