

회귀해석 프로젝트

서울시 공공자전거 따릉이의 이용자 수 예측

동국대학교 통계학과

2018110482 김 언 지 ,

2018110479 김 은 지

2 0 2 0

목 차

제1장 서론	1
제1절 주제 선정 및 프로젝트의 목적	1
제2장 본론	2
제1절 데이터 소개	2
1. 데이터 소개 및 변수 소개	2
제2절 데이터 전처리 및 분석	3
1. 변수 추가	3
2. 상관계수 크롤링	4
3. 공원 개수 counting	6
4. 한강공원 유/무 및 한강공원 종류 변수 수집	6
5. 결측치/이상치 처리	7
제3절 모델링	8
1. 다중회귀모델 생성	8
2. 회귀모델 데이터 예측	13
제3장 결론	15
제1절 분석결과 정리 및 기대효과	15

참 고 문 헌

<표 목 차>

<표 1> 서울특별시 공공자전거 대여소 정보(19,12,9).xlsx
<표 2> 서울특별시 공공자전거 대여정보_201901.csv
<표 3> 서울특별시 공공자전거 이용정보(월별)_201901_201906.xlsx
<표 4> final.csv

[그림목차]

[그림 1, 2] 따릉이 추가배치 지연에 대한 시민의견과 서울특별시 따릉이 측의 답변
[그림 3, 4] 소상공인 상권정보사이트 상권분석시스템-1단계 지역선택,2단계 영역선택
[그림 5, 6] 소상공인 상권정보사이트 상권분석시스템-3단계 업종선택 및 분석하기
[그림 7] 상권분석 결과 화면(상권평가 탭(좌)과 지역분석 탭(우))
[그림 8,9] 한강공원 지도(공원 범위 확인)
[그림 10, 11] 주소 검색이 안 되는 경우 - 크롤링 시 오류 발생 (NA 처리됨)
[그림 12] 계절별 이용자 수 barplot
[그림 13] 계절별 이용자 수 boxplot
[그림 14] 이상치 제거 후 계절별 이용자수 boxplot
[그림 15] 최종 데이터의 예측변수들간, 예측변수-반응변수간의 scatter plot 확인
[그림 16] 최종 데이터의 예측변수들간, 예측변수-반응변수간의 상관관계 plot

- [그림 17] 문자형 변수만 지우고 바로 돌린 full model 결과창
- [그림 18] 'n_hanriverpark','oneDay_mem','oneDay_cust','regular' 제거모델결과
- [그림 19] 위 모델에서 다중공선성 높은 'total_store'변수 제거한 모델 결과 창
- [그림 20] 최종 모델의 회귀식을 가지는 model
- [그림 21] 이상치 제거 모델 plot
- [그림 22] 이상치제거모델의 표준화잔차히스토그램
- [그림 23] 최종 회귀 모델 plot
- [그림 24] 최종회귀모델의 표준화잔차히스토그램
- [그림 25] 최종회귀모델 QQ-plot
- [그림 26] 잔차 대 순서 플랏
- [그림 27] 최종회귀모델 결과창
- [그림 28] \hat{y} vs y plot
- [그림 29] 10-folds cross validation 결과
- [그림 30] 모델의 성능 분산(오차)

제1장 서론

제1절 주제 선정 및 프로젝트의 목적

서울특별시 2015년 10월부터 ‘따릉이’라고 불리는 무인 공공 자전거 대여 서비스를 시작했다. 직장인들의 출·퇴근 시 이용을 도와 자전거 교통 분담률을 높이고 친환경 자전거 도시를 만든다는 일환으로 실시되었다. 따릉이는 지하철 출입구·버스정류장·관공서·학교 등 접근이 쉬운 통행 장소 등에 대여소가 설치돼 있어 이용이 편리하다. 대여소에서는 따릉이의 대여와 반납이 무인으로 이뤄지며, 이용자들은 장소에 구애받지 않고 대여소가 설치된 곳이면 어디에서나 자전거를 대여하고 반납할 수 있다. 특히, 한강공원이나 월드컵 공원 등 자전거를 타기 좋은 공원의 대여소나, 회사밀집지역, 학교밀집지역의 대여소는 대여자수가 많아 배치되어있는 자전거가 없어서 많은 이용자들이 이용하고 싶어도 이용을 할 수 없는 경우가 발생한다. 또는 망월한강공원에서 망월역으로 가는 경로에 있는 152. 마포구민체육센터앞(구) 대여소는 반납자수가 많아 거치대수보다 많은 자전거가 배치되어있다. 쉬운 접근성과 낮은 이용료, 높은 인지도로 따릉이의 이용자수는 해마다 큰 증가폭을 보인다. 하지만 이용자수의 증가폭에 비해 따릉이 측에서는 따릉이 대여소의 자전거 추가배치나 거치대 추가 같은 요청에 대한 피드백이 느린 편이다[그림 2,3].

시민의견수렴

추가배치가 몇시간씩 되지않고 방치되는것에 관하여
2020.04.21 ag****

자전거 추가배치요청을 따로 하지않을 시 몇시간씩 배치가 안되거나
배치요청을해도 시간이 정해져있지않아 무한정 기다려야하는 경우가 많습니다. 저도 지금배치요청을 한지 40분넘도록 배치가 안되고있어요.
요청시 언제 배치가되는지 알수있으면 좋겠고 배치시 알림기능이 만들어지면 좋겠습니다.
또한 비어있는 거치대는 빨라야 아니라도 몇시간 안에는 그래도 배치를 하는 규정이 있으면 좋겠어요.

또 50대이상 어르신들이 자주 따릉이 어떻게타냐고 묻는데 어르신들도 쉽게 대여할수있도록 티머니같은 카드를 만들거나해서 어르신들 이용이 수월하도록해주세요

시민의견수렴

따릉이 재배치 시간
2020.04.20 vc*****

따릉이 재배치하는 시간이 어떻게 되나요?? 저희 지역은 아침이 아니면 이용하기가 힘들 거 같은데요?? 배치 요청 눌러도 언제 재배치 해주는지 모르니 이용하기가 힘듭니다. 일이 있는데 자전거 올 때까지 마냥 기다릴 수도 없고, 주변 대여소가 전부 0이예요 괜히 고장난 자전거 고쳐서 이용해보려다가 손만 다치고, 고장신고 하면 이용 가능한 자전거에 안타까 하거나 아닐까요? 고장신고 하러 했더니 이미 접수된 자전거래요.

안녕하세요.
서울공공자전거 따릉이입니다.
먼저 따릉이 이용에 시민님께 불편을 드려 대단히 죄송합니다.
원활한 운영을 위하여 대여소의 자전거 회수 및 거치를 지속적으로 수행하고 있지만, '따릉이' 이용의 증가와 지역별 특성, 특정 시간대에 따른 자전거 유통현상이 과도하게 나타나 분배의 어려움을 겪고 있습니다.
또한, 코로나바이러스-19 관련 예방을 우선적으로 진행하기 위하여, 1500여개의 대여소를 매일 전수 소독하고 있으며 정비 전후의 자전거를 추가소독하고, 전 대여소 내 손세정제 비치 및 관리 등을 실시하고 있습니다.
이에 따라 자전거의 원활한 배치에 어려움이 있을 수 있음을 너그럽게 이해하여주시기 바랍니다.
이에 대하여 분배팀이 더 신경을 쓰며 관리 하도록 하겠습니다.
의견주신 자전거대여소가 어디인지 말씀해주시면 지역 담당 분배팀에 전달하여 이용 불편을 줄이도록 내용 전달하겠습니다.
앞으로도 많은 이용과 관심 부탁드립니다.
따릉이 이용과 관심에 감사드립니다.
오늘도 좋은 하루 보내시길 바랍니다.
감사합니다.

[그림 1, 2] 따릉이 추가배치 지연에 대한 시민의견과 서울특별시 따릉이 측의 답변¹⁾

따라서 본 조는 주변 상권의 특징과 주거인구수, 직장인구수, 유동인구수, 학교의 수, 이용자의 나이, 대여정보가 따릉이의 이용자수에 영향을 주는지 다중선형회귀모델을 통해 확인해보고, 상권의 특징, 주거인구, 직장인구, 유동인구수, 학교 유무, 계절변수, 한강유무변수, 한강종류변수, 공원유무변수, 지역변수 등을 토대로 따릉이 이용자수를 예측해보고자 한다. 예측 결과를 통해, 자전거 추가배치나 거치대 추가 등의 대여소 관리를 더 계획적이고 유동적이고 효율적으로 할 수 있을 것이라고 기대한다.

1) (출처: <https://www.bikeseoul.com/main.do>) 현재 서울시 따릉이 사이트에선 시민의견수렴 탭을 확인 할 수 없다. (캡처 일자 2020.04.22.)

제2장 본론

제1절 데이터 소개 및 변수 소개

따릉이 이용자수, 반납수에 대한 대여수의 비율을 예측하기 위해 ‘서울 열린데이터 광장’을 통해 서울시 공공자전거에 대한 여러 가지 데이터를 수집하였다. 필요한 독립변수들과 종속변수를 수집하기 위해 사용한 데이터의 내용은 다음과 같다.

1. 대여소 정보

‘서울특별시 공공자전거 대여소 정보(19,12,9).xlsx’를 통하여 총 대여소의 정보에 대해 알 수 있었다.

대여소_구	대여소ID	대여소명	대여소 주소	위도	경도	기준 시작일자	거치대 수
마포구	101	101.(구) 합정동주민센터	서울특별시 마포구 동교로 8길 58	37.54956	126.90575	2015-09-06 23:40	5
...
합계		1,540					19,545

<표 1> 서울특별시 공공자전거 대여소 정보(19,12,9).xlsx

이 데이터를 통해 총 1,540개의 대여소 정보를 얻을 수 있었다. 이를 통해 대여소_구, 대여소 ID, 대여소명, 대여소주소, 위도, 경도, 기준시작일자, 거치대수를 알 수 있었다.

2. 월별 대여 정보

‘서울특별시 공공자전거 대여정보_201901.csv’를 통하여 1월의 대여정보에 대해 알 수 있었다.

자전거 번호	대여 일시	대여 대여소번호	...	반납일시	반납 대여소번호	...	이용 시간	이용 거리
SPB-1 0957	2019-01-01 12:02:16	1408	...	2019-01-01 12:07:07	1433	...	4	1020
...

<표 2> 서울특별시 공공자전거 대여정보_201901.csv

이 데이터를 통해 대여일시가 2019년 1월의 대여정소에 대해 알 수 있었다. 한 행은 하나의 이용건수로 이루어져있으며 자전거번호, 대여일시, 대여대여소번호, 대여대여소명, 대여거치대순번, 반납일시, 반납대여소번호, 반납대여소명, 반납거치대순번, 이용시간, 이용거리에 대한 정보가 나타나있다.

이와 같은 (월별로 나뉜)데이터로 2월, 3월, 4월~ 10월까지의 각 월별 대여정소에 대해 정보를 얻을 수 있었다.

3. 대여소의 이용자 정보

‘서울특별시 공공자전거 이용정보(월별)_201901_201906.xlsx’를 통해 2019년 1월부터 6월까지의 이용자 정보를 알 수 있었다.

대여일자	대여소번호	...	대여구분코드	성별	연령대 코드	이용건수	...
2019-01-01	3	...	일일(회원)	M	AGE_003	12	...
...

<표 3> 서울특별시 공공자전거 이용정보(월별)_201901_201906.xlsx

이 데이터는 한 행이 대여월, 대여소번호, 대여구분코드, 성별, 연령대의 정보가 같은 이용자로 이루어져있고 이용건수를 통해 정보가 같은 이용자 명수를 표시해져있다. 정보가 같은 이용자가 각 대여소에서 대여를 했을 때, 대여일자, 대여소번호, 대여소명, 대여구분코드(일일(회원), 일일(비회원), 정기, 단체), 성별(M,F), 연령대코드(AGE_001, AGE_002, AGE_003, ... ,AGE_008 : 10대, 20대, 30대, ... , 80대), 이용건수, 운동량, 탄소량, 이동거리(M), 이동시간(분)의 정보를 알 수 있다.

제2절 데이터 전처리 및 분석

앞서 2장.1절에서 소개한 데이터를 통해 독립변수들과 종속변수를 생성하여 분석 전 최종 데이터('bicycle.csv')를 생성하였다. 필요한 정보(독립변수들과 종속변수를 열로 생성)를 얻어 생성한 데이터 'bicycle.csv'는 한 행이 한 대여소 정보(총 1,540대여소)가 되고 월별(총 11월)로 1,540*11 = 16,940행이 생성했다.

bicycle데이터에서 결측치/이상치 처리한 최종 데이터는 '최종final이상치제거(Turkey Fences).csv' 로 저장하였다.

최종 데이터에 들어가는 변수에는,

1) 대여소 정보 변수

: id(대여소ID), name(대여소명), area5(대여소주소_권역별), latitude(해당 대여소의 위도), longitude(해당 대여소의 경도), number_of_holders(거치대수), season(이용 날짜-계절)

2) 종속변수와 관련된 변수

: rent(대여수), retrun(반납수)

3) 대여소 근처의 상권 정보 변수

: total_store(전체 업소 수), food_num(음식점 수), live_num(주거 인구 수), work_num(직장 인구 수), move_num(유동 인구 수), school_YN(학교 유무), traffic(지하철 수+버스정류장 수), park_YN(공원 유무), in_hanriverpark(주변한강유무), n_hanriverpark(주변한강종류-이름)

4) 이용자 정보 변수

: oneDay_mem(일일(회원) 이용자 수), oneDay_cust(일일(비회원) 이용자 수), regular(정기 이용자 수), group(단체 이용자 수), AGE_001(10대 이용자 수), AGE_002(20대 이용자 수), AGE_003(30대 이용자 수), AGE_004(40대 이용자 수), AGE_005(50대 이용자 수), AGE_006(60대 이용자 수), AGE_007(70대 이용자 수), AGE_008(80대 이용자 수)

id	name	season	...	rent	return	oneDay_mem	...	total_store	...	AGE_001	...
101	101. (구)합정동 주민센터	겨울	...	239	241	22	...	12	...	0	...
...
664	서울시립대 대학본부	가을	...	126	124	13	...	0	...	23	...

<표 4> final.csv

1. 변수 추가

각 대여소의 월별 rent(대여수), return(반납수)를 수집하기 위해서 ‘대여소 정보’ 데이터와 ‘월별 대여 정보’ 데이터를 사용하였다.

대여수와 반납수를 수집하기에 앞서, 각 월별로 이용된 대여소를 살펴보았다. 해당 월에 이용된 대여소 중, 최종 대여소 정보에 없는 대여소가 존재했다. 1월부터 10월까지 총 30개의 대여소였다. 이들은 최종적으로 직원들이 사용하는 대여소(11개), 폐쇄된 대여소(19개)로 판단하여 주제와는 맞지 않는 대여소였기에 ‘월별 대여 정보’에서 제거하였다.

그리고 월별 대여수와 반납수를 파악하였다. 대여수는 대여소 정보에 있는 대여소번호와 월별 대여 정보의 대여대여소번호가 일치하면 해당 대여소의 대여수에 +1을 하는 방식으로 이루어졌다. 반납수도 대여수와 같은 방식으로, 대여소 정보에 있는 대여소번호와 월별 대여 정보의 반납대여소번호가 일치하면 해당 대여소의 반납수에 +1을 하는 방식으로 이루어졌다.


이용정보와 관련한 대여구분 변수 총 4개와 나이(10대부터 80대까지) 변수 총 8개는 ‘대여소 정보’ 데이터와 ‘월별 이용자 정보’ 데이터를 사용하여 수집하였다. 분석의 목적이 ‘예측’에 있기에 분석에 유의미하기 위해서 바로 이전 월의 연령대별, 대여구분별로 연령대, 대여구분 값을 채워 넣었다. 대여구분 변수 oneDay_mem, oneDay_cust, regular, group은 대여소 정보에 있는 대여소번호와 이전의 월별 이용자 정보의 대여소번호가 일치할 때, 대여구분코드가 일일(회원)이면 oneDay_mem 변수에 + 이용건수, 일일(비회원)이면 oneDay_cust 변수에 + 이용건수, 정기이면 regular 변수에 + 이용건수, 단체이면 group 변수에 + 이용건수를 하면서 대여소별 대여구분별 이용자 수를 수집하였다. 예를 들면, 7월 이용자수를 예측하기위해 6월의 이용자수가 대여구분별로 몇 명 있는지 수집하였다.

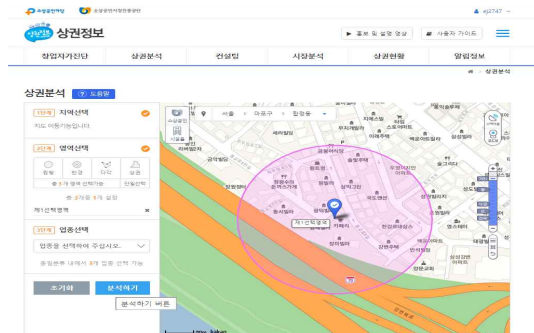
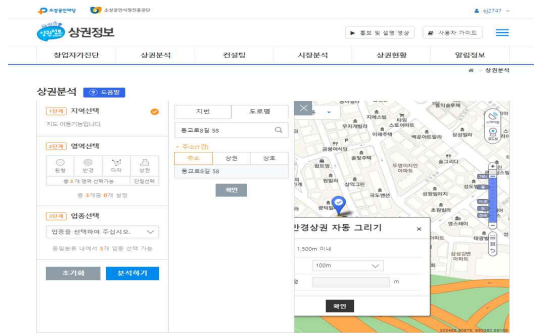
나이 변수도 대여구분 변수와 같은 방식으로 대여소 정보에 있는 대여소번호와 이전의 월별 이용자 정보의 대여소번호가 일치할 때, 나이에 따라 각 나이 변수에 + 이용건수를 하는 방식으로 대여소별 연령별 이용자 수를 수집하였다.

2. 상관계이터 크롤링

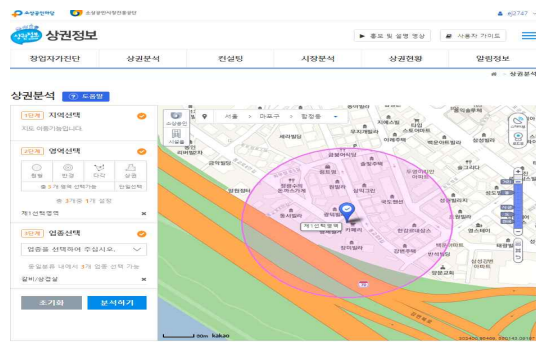
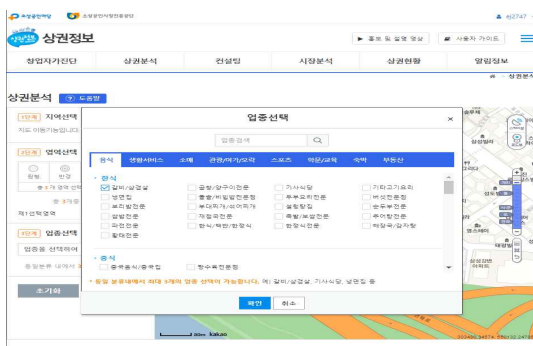
1540개의 따릉이 대여소 주변 상관계이터는 소상공인 상권정보 사이트 (<http://sg.sbiz.or.kr/main.sg#/main>) 에서 R selenium 패키지를 이용해 크롤링을 통해 가져왔다.

대여소의 주소는 ‘서울특별시 공공자전거 대여소 정보(19.12.9).xlsx’ 에서 가져왔으며, “서울특별시 ○○구 ○○동 ○○-○○” 또는 “서울특별시 ○○구 ○○로 ○○” 와 같이 “서울특별시”와 “○○구”가 포함된 주소는 해당 사이트에서 주소 검색이 되지 않아 “○○동 ○○-○○” 또는 ‘○○로 ○○’와 같이 주소를 분리해 새로운 주소 벡터(doroAddr)를 만들었다. 이 주소벡터에는 대여소 1540개의 주소가 들어있으며 지번 주소 및 도로명 주소가 지역선택 란에 자동으로 입력되게 하였다.

새로 만든 주소벡터(doroAddr)로 주소검색이 완료되면(ex. 동교로8길 58), 해당 주소에 핀() 이 찍힌다. 그 핀의 xpath를 읽어와 2단계 영역선택 탭에서 반경을 선택해 핀의 xpath를 누르게 하면 자동으로 지정한 반경이 그려진다. 해당 사이트의 반경 기본 값은 100m이고, 본 조가 생각하기에 상점 또는 회사(학교) 에서(으로) 걸어가기에 적합한 반경이 100m라고 판단하여 반경을 100m로 설정하여 분석했다[그림 6].



[그림 3, 4] 소상공인 상권정보사이트 상권분석시스템-1단계 지역선택, 2단계 영역선택, 3단계 업종선택 탭은 아래 그림과 같이 음식, 생활서비스, 소매, 관광/여가/오락, 스포츠, 학문/교육, 숙박, 부동산으로 분류되며 각각 하위분류를 가진다. 하지만, 어떤 업종을 택하든, 또는 업종을 1개만 택하든, 3개를 택하든 결과화면에서 가져올 정보는 동일하므로 가장 앞에 있는 갈비/삼겹살 업종을 선택해 분석하였다.



[그림 5, 6] 소상공인 상권정보사이트 상권분석시스템-3단계 업종선택 및 분석하기 크롤링을 통해 3단계 업종선택까지 완료하여 분석하기 버튼을 누르면 해당 주소의 반경100m의 상권분석 보고서 결과화면이 뜬다. 우리는 6개의 탭 중 상권평가 탭에서 전체 업소수, 음식 업소수, 주거 인구수, 직장 인구수, 유동 인구수를 가져오고, 지역분석 탭에서 학교 수와 지하철역 수, 버스정류장 수를 가져왔다. 이 때 학교 수보다는 학교 유무가 중요하다고 판단 되어 (학교 수)>0이면 '1' (학교 수)=0이면 '0'의 값을 가지는 학교유무변수(school_YN)을 만들었다. 그리고 지하철역 수와 버스정류장 수는 서로 비슷한 성질을 지니는 변수이므로 따로 해석하기보다는 합계를 이용하기로 하였다. 이는 traffic변수로 이용했다.



[그림 7] 상권분석 결과 화면(상권평가 탭(좌)과 지역분석 탭(우))

3. 공원 개수 counting

상권 정보 이외에도 대여소의 특성으로 '공원 정보 변수'를 추가하였다. 서울시 공공데이터포

텔 사이트에서 '서울시 주요 공원현황.csv'을 이용하여 서울의 1382개의 주요 공원의 위도및경도를 가져왔다. 1382개 주요 공원을 이용한 이유는 집 앞의 작은 규모의 공원보다는 규모가 어느 정도 되는 공원이 자전거를 탈 수 있는 환경을 갖추었을거라고 생각했기 때문이다. 공원의 개수는 공원과 대여소의 위도 경도를 소수 둘째자리까지 반올림하여 같은 그룹에 속하면 공원의 개수를 +1하는 방식으로 진행했다. 예를 들어 한 공원의 위도-경도가 "37.55 - 126.91"이고 대여소의 위도-경도가 "37.55 - 126.91"이면 park_num에 +1을 하였다. 이렇게 생성된 park_num이 park_num>0을 만족하면 park_YN에 '1', park_YN에 '0'을 저장했다.

4. 한강공원 유/무 및 한강공원 종류 변수 수집

따릉이 대여소가 한강공원 안에 속하는지, 속한다면 어떤 한강공원 안에 속하는지를 각각 '한강공원 유무 변수(in_hanriverpark)'와 '한강공원 종류 변수(n_hanriverpark)'로 수집하였다. 서울의 한강공원은 총 11개로 (광나루, 잠실, 독섬, 잠원, 이촌, 반포, 망원, 여의도, 난지, 강서, 양화) 한강공원이 있다. 한강공원의 범위는 '서울특별시 한강사업본부' 사이트를 기준으로 하였다.

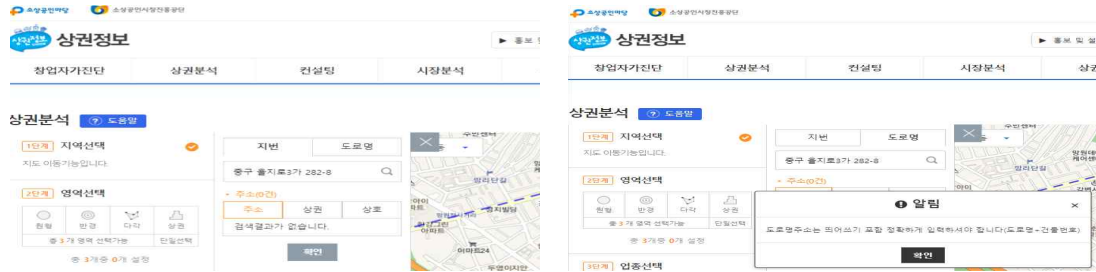


[그림 8.9] 한강공원 지도(공원 범위 확인)

정해진 범위를 따릉이 어플'을 활용해 같은 범위를 파란 선으로 표시하고, 그 안에 속하는 따릉이 대여소 보니 한강 공원 안에 따릉이 대여소가 많을 것이라고 예상했던 것과 달리 매우 적었다. '여의도 한강공원'에 대여소가 31개로 제일 많았고, '강서 한강공원'이 0개로 제일 작았다. 그 이유는 '서울자전거 따릉이'사이트의 건의사항에서 알 수 있었다. 안전을 위한 것과 한강관리 사업소에 많은 금액을 내고 자전거 임대업으로 생계를 유지하는 분들이 있기에 한강공원 안에는 따릉이 대여소를 거의 설치하지 않아 한강공원 안에는 1540개의 대여소 중 총 80개밖에 없어 직접 한강공원별 대여소 번호를 수집할 수 있었다. 이를 엑셀파일로 저장한 후, 해당하는 대여소 번호 행의 한강유무변수에는 '1', 한강종류변수에 '해당 한강공원 이름'으로 저장해 2개의 범주형 변수를 생성하였다.

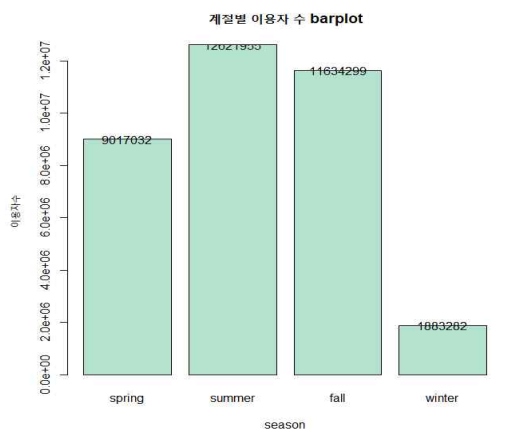
5. 결측치/이상치 처리

위의 2. 상관계이터 크롤링 과정에서 지번으로도 도로명으로도 검색이 안 되는 주소들이 약 10%가량 생겼다. 이는 해당 대여소의 위도·경도를 네이버 지도나 구글 지도에 검색하여 검색 가능한 주소를 가져와 1단계 지역선택 란에 입력하여 검색하는 방법으로 해결할 수 있었지만, 시간이 오래 소요되므로 시간관계 상 실행하지 못하여 1단계 지역선택 란을 생략하고 2단계 영역선택 란에서 직접 대여소 위치에 핀을 찍었다. 그 후의 과정은 동일하다. 반경 100m영역을 만든 다음, 3단계 업종선택 란에서 갈비/삼겹살 업종을 선택하여, 결과화면에서 전체 업소수, 음식 업소수, 주거 인구수, 직장 인구수, 유동 인구수, 학교 수, 지하철역 수, 버스정류장 수를 가져왔다.

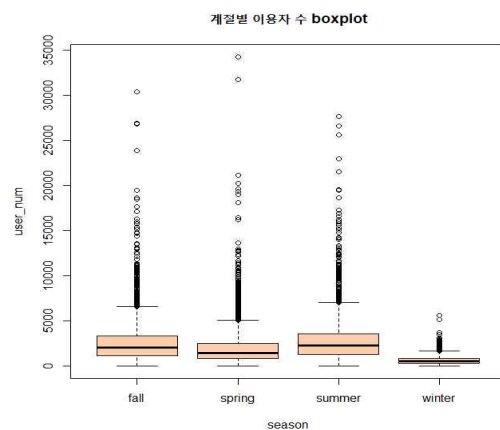


[그림 10, 11] 주소 검색이 안 되는 경우 - 크롤링 시 오류 발생 (NA 처리됨)

또한 이용자수(대여수+반납수)가 0인 행은 '서울특별시 공공자전거 대여소 정보(19.12.9).xlsx'에서 확인해본 결과, 해당 월에 존재하지 않았던 대여소이므로(그 이후에 생긴 신규 대여소, 또는 폐쇄된 대여소) 모델 생성에 필요는 없지만 모델에 영향을 줄 수 있을 것이라 생각하여 행삭제를 진행하였다.



[그림 12] 계절별 이용자 수 barplot

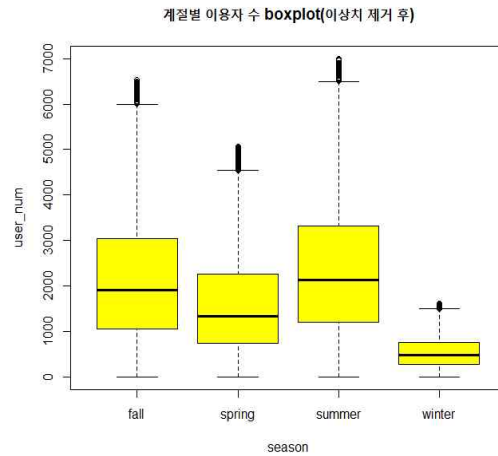


[그림 13] 계절별 이용자 수 boxplot

[그림13]을 통해 계절별로 따릉이 이용자 수가 차이를 보이는 것을 알 수 있었다. 이는 자전거를 타기에 힘든 날씨(겨울, 가을) 보다는 자전거를 타기에 적합한 날씨인 (봄,여름)에 이용자 수가 많은 것을 알 수 있었다.

[그림14]를 통해 월별 이용자 수의 boxplot을 확인하였다. 이를 통해 계절별로 quantile값이 차이 나는 것과 이상치가 꽤 있음을 확인할 수 있다. 이상치는 회귀모델생성에 있어서 큰 문제점을 야기할 수 있다. 따라서 전처리 과정에서 최종 데이터의 이상치를 제거하였다. 이상치 제거 방법으로는 'Z-score', 'Turekey Fences' 방법을 사용하였으며, 둘 중 Turkey Fences 방법으로 제거한 데이터를 최종 데이터로 이용하였다. Turkey Fences 방법으로 $Q3+1.5*IQR$ 이상인 행과 $2)Q1-1.5*IQR$ 대신 user_num이 0인 행을 제거하였더니 데이터의 95.23%가 남았다. 두 방법 중 Turkey Fences 방법을 이용한 이유는 제거된 이상치의 개수는 비슷하지만, 이 방법이 이상치 제거한 데이터를 boxplot 을 통해 이상치가 제거된 후의 변화를 시각적으로 확인할 수 있기 때문이다.

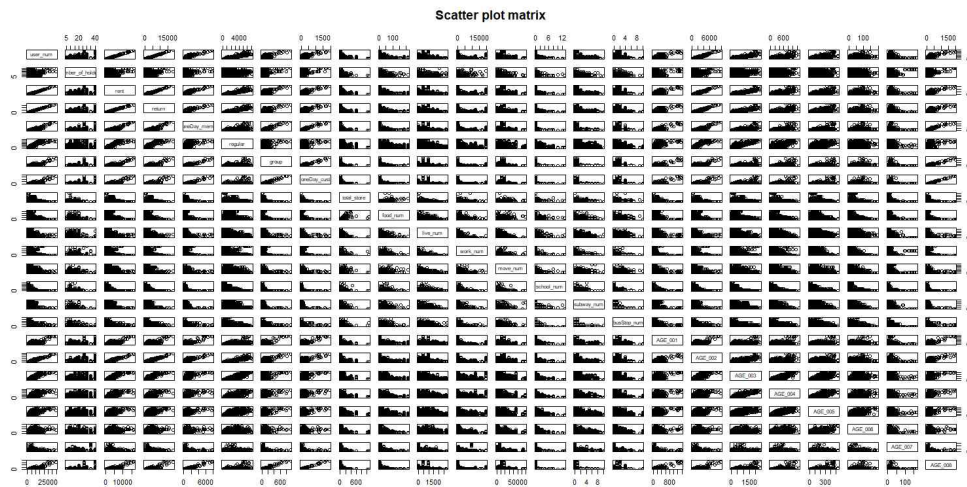
2) user_num은 이용자 수 이므로 항상 0보다 커야 한다. 따라서 $Q1-1.5IQR$ 로 제거하기 보다는 0인 데이터를 이상치로 판단하고 제거하였다.



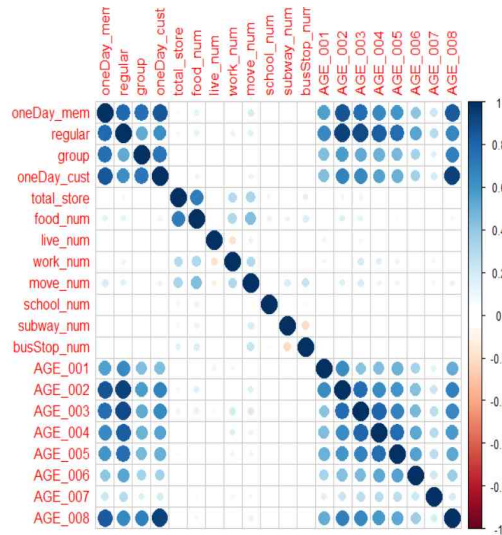
[그림 14] 이상치 제거 후 계절별 이용자수 boxplot

제3절 모델링

1. 다중회귀모델(Multiple Linear Regression Model) 생성



[그림 15] 최종 데이터의 예측변수들간, 예측변수-반응변수간의 scatter plot 확인



[그림 16] 최종 데이터의 예측변수들간, 예측변수-반응변수간의 상관관계 plot

최종 데이터를 가지고 scatter plot을 그려보았다. 변수개수가 많아 scatter plot을 선형관계를 확인하기 어려워 corrplot을 그려 상관관계가 높은 변수들을 확인하였다(범주형 변수도 범주를 숫자로 바꿔 상관관계 확인). 최종 데이터를 넣고 돌린 full model의 결과는 다음과 같았다.

```
Call:
lm(formula = user_num ~ ., data = df2_1)

Residuals:
    Min       1Q   Median       3Q      Max
-5466.8  -209.1   -10.4   162.5   3704.5

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.236e+01  1.919e+01  -1.686  0.091828 .
area5동남권  1.463e+01  1.361e+01   1.075  0.282546 .
area5동북권  9.487e+01  1.377e+01  6.890  5.80e-12 ***
area5서남권  8.256e+01  1.363e+01  6.057  1.42e-09 ***
area5서북권  1.596e+00  1.514e+01  0.105  0.916037
number_of_holders -3.865e+00  7.829e-01  -4.937  8.04e-07 ***
oneDay_mem    7.673e-01  9.147e-02   8.388  < 2e-16 ***
regular      -2.228e-02  3.263e-02  -0.683  0.494617
group         4.140e-01  4.995e-01   0.829  0.407234
oneDay_cust  -1.487e+01  1.045e+00 -14.239  < 2e-16 ***
total_store  -1.041e-01  8.252e-02  -1.262  0.207096
food_num     1.694e+00  2.435e-01  6.957  3.62e-12 ***
live_num     -2.030e-03  7.109e-03  -0.286  0.775194
work_num     -7.162e-03  2.183e-03  -3.281  0.001037 **
move_num     2.947e-03  6.083e-04  4.844  1.29e-06 ***
AGE_001      1.114e+00  1.332e-01  8.367  < 2e-16 ***
AGE_002      1.826e+00  3.977e-02  45.924  < 2e-16 ***
AGE_003      2.032e+00  6.186e-02  32.842  < 2e-16 ***
AGE_004      2.270e+00  8.833e-02  25.701  < 2e-16 ***
AGE_005      2.346e+00  1.327e-01  17.677  < 2e-16 ***
AGE_006      3.907e+00  2.829e-01  13.810  < 2e-16 ***
AGE_007      9.490e-02  5.662e-01  0.168  0.866882
AGE_008      1.270e+01  9.498e-01  13.375  < 2e-16 ***

in_hanriverpark1 -1.652e+02  6.375e+01 -2.591  0.009581 **
n_hanriverpark관나루 4.042e+02  9.735e+01  4.152  3.31e-05 ***
n_hanriverpark난지  1.678e+03  2.404e+02  6.982  3.04e-12 ***
n_hanriverpark독섬  3.482e+02  7.782e+01  4.475  7.71e-06 ***
n_hanriverpark왕원  2.910e+02  9.771e+01  2.978  0.002905 **
n_hanriverpark반포  3.944e+02  1.560e+02  2.529  0.011442 *
n_hanriverpark암화  2.800e+02  8.294e+01  3.376  0.000737 ***
n_hanriverpark여의도 5.094e+01  6.899e+01  0.738  0.460298
n_hanriverpark이촌  2.360e+02  8.741e+01  2.700  0.006935 **
n_hanriverpark잠실  3.084e+02  7.667e+01  4.022  5.80e-05 ***
n_hanriverpark잠원  NA NA NA NA
seasonspring  5.738e+02  1.210e+01  47.416  < 2e-16 ***
seasonsummer -1.185e+02  1.199e+01 -9.880  < 2e-16 ***
seasonwinter -4.832e+01  1.308e+01 -3.693  0.000222 ***
park_YN1     -1.073e+01  8.178e+00 -1.312  0.189495
school_YN1   -2.901e-01  8.865e+00 -0.033  0.973892
traffic      -4.250e+00  2.179e+00 -1.951  0.051107 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 400.5 on 14442 degrees of freedom
Multiple R-squared:  0.9219, Adjusted R-squared:  0.9217
F-statistic: 4488 on 38 and 14442 DF, p-value: < 2.2e-16
```

[그림 17] 문자형 변수만 지우고 바로 돌린 full model 결과창

ols 패키지를 이용해 full 모델의 다중공선성을 확인한 결과 vif >10인 변수들은 'oneDay_mem', 'regular', 'oneDay_cust', 'AGE_002', 'AGE_008', 'in_hanriverpark', 'n_hanriverpark' 로, 특히 'in_hanriverpark'와 'n_hanriverpark'는 다중공선성으로 inf값을 가진다. 실제로 한강종류와 한강유무변수의 correlation을 확인했더니 약 0.9383으로 매우 높은 상관관계를 가짐을 알 수 있다. 따라서 두 변수 중 기본적으로 한강유무변수가 y변수와 유의하고, 한강종류변수는 너무 많은 범주를 가져 해석이 어렵다는 단점이 있어 둘 중 '한강유무변수(in_hanriverpark)'를 선택했다. 그리고 해석이 어려운 'oneDay_mem', 'regular', 'oneDay_cust' 보다는 해석이 쉬운 'AGE_002', 'AGE_008'을 선택하였다. 1)변수선택과정에서 최적의 변수개수 확인플랏을 고려하면서 진행했다.

full model의 다중공선성이 높은 변수 4개를 제거하고 다시 회귀를 돌린 결과이다.

```
Call:
lm(formula = user_num ~ ., data = df3)

Residuals:
    Min       1Q   Median       3Q      Max
-5169.8  -211.1   -11.3   165.4   3737.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.086e+01  1.873e+01  1.648  0.09947 .
area5동남권  3.044e+01  1.337e+01  2.277  0.02281 *
area5동북권  9.338e+01  1.361e+01  6.859  7.23e-12 ***
area5서남권  7.703e+01  1.324e+01  5.817  6.13e-09 ***
area5서북권  9.541e+00  1.496e+01  0.638  0.52362
number_of_holders -4.768e+00  7.847e-01  -6.076  1.26e-09 ***
group         1.230e+00  4.421e-01   2.782  0.00541 **
total_store  -1.008e-01  8.340e-02  -1.209  0.22667
food_num     1.541e+00  2.455e-01  6.279  3.50e-10 ***
live_num     3.231e-03  7.102e-03   0.455  0.64912
work_num     -9.015e-03  2.177e-03  -4.141  3.48e-05 ***
move_num     3.488e-03  6.110e-04  5.724  1.06e-08 ***
AGE_001      1.183e+00  1.297e-01  9.122  < 2e-16 ***
AGE_002      1.927e+00  2.250e-02  85.640  < 2e-16 ***
AGE_003      2.084e+00  5.313e-02  39.233  < 2e-16 ***
AGE_004      2.213e+00  8.297e-02  26.673  < 2e-16 ***
AGE_005      2.346e+00  1.300e-01  18.046  < 2e-16 ***
AGE_006      3.699e+00  2.823e-01  13.105  < 2e-16 ***
AGE_007      -5.356e-01  5.683e-01  -0.942  0.34600
AGE_008      1.504e+00  3.122e-01  4.816  1.48e-06 ***
in_hanriverpark1  3.882e+01  1.716e+01  2.262  0.02368 *
seasonspring  5.132e+02  1.108e+01  46.304  < 2e-16 ***
seasonsummer -1.953e+02  1.018e+01 -19.183  < 2e-16 ***
seasonwinter -1.077e+02  1.238e+01 -8.699  < 2e-16 ***
park_YN1     -1.933e+01  8.197e+00 -2.359  0.01834 *
school_YN1   3.275e+00  8.922e+00  0.367  0.71357
traffic      -3.965e+00  2.200e+00 -1.803  0.07146 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 405.4 on 14454 degrees of freedom
Multiple R-squared:  0.92, Adjusted R-squared:  0.9198
F-statistic: 6389 on 26 and 14454 DF, p-value: < 2.2e-16
```

[그림 18] 'n_hanriverpark', 'oneDay_mem', 'oneDay_cust', 'regular' 제거모델결과

그리고 변수들 간 교호작용도 고려해봤다. 먼저 모든 가능한 교호작용을 넣은 모델을 생성하고 그 모델에서 t-test p-value가 유의한 교호작용변수를 선택하여 기존 모델식에 추가하였다. 유의한 교호작용 변수들로는

area5*AGE_003 + number_of_holders*total_store+ number_of_holders*food_num + number_of_holders*work_num + number_of_holders*AGE_005 + group*work_num + group*AGE_004 + group*AGE_005 + group*AGE_008 + group*traffic+ total_store*food_num+ total_store*move_num + food_num:move_num + live_num*move_num 이 있었다. 하지만 교호작용을 넣고 돌린 모델에서 다중공선성이 다수 발견되었다. 다중공선성이 높은 변수들은 total_store, food_num, AGE_003, number_of_holders:total_store, number_of_holders:food_num, group:AGE_004, group:AGE_005, total_store:move_num, food_num:move_num로, 먼저 total_store와 food_num과 다중공선성이 높은 변수들이 많기 때문에 total_store와 food_num 중 더 다중공선성이 높은 total_store를 제거하기로 했다.

<pre>Call: lm(formula = user_num ~ . + area5 * AGE_003 + number_of_holders * food_num + number_of_holders * work_num + group * AGE_004 + group * AGE_005 + group * AGE_008 + group * traffic + food_num * move_num + live_num * move_num, data = df3.1)</pre>					<pre>area5동남권:AGE_003 1.17e-01 7.461e-02 1.577 0.114801 area5동북권:AGE_003 3.614e-01 7.771e-02 4.651 3.33e-06 *** area5서남권:AGE_003 2.657e-01 7.207e-02 3.687 0.000228 *** area5서북권:AGE_003 3.118e-01 8.283e-02 3.764 0.000168 *** number_of_holders:food_num -1.571e-03 3.750e-02 -0.042 0.966577 number_of_holders:work_num -4.454e-04 3.365e-04 -1.324 0.185687 group:AGE_004 -2.362e-02 3.126e-03 -7.556 4.39e-14 *** group:AGE_005 -2.212e-02 5.945e-03 -3.721 0.000199 *** group:AGE_008 -3.891e-02 3.502e-03 -11.109 < 2e-16 *** group:traffic -7.186e-01 1.520e-01 -4.729 2.28e-06 *** food_num:move_num -6.141e-05 1.473e-05 -4.169 3.08e-05 *** live_num:move_num -2.687e-06 1.284e-06 -2.092 0.036440 *</pre>				
<pre>Residuals: Min 1Q Median 3Q Max -5034.4 -212.8 -11.3 166.2 3850.1</pre>					<pre>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</pre>				
<pre>Coefficients: (Intercept) -8.266e+01 2.408e+01 -3.433 0.000599 *** area5동남권 2.151e+01 2.026e+01 1.062 0.288318 area5동북권 5.283e+01 1.996e+01 2.647 0.008139 ** area5서남권 4.788e+01 1.989e+01 2.407 0.016079 * area5서북권 -2.770e+01 2.242e+01 -1.236 0.216581 number_of_holders -4.610e+00 9.430e-01 -4.888 1.03e-06 *** group 1.483e+01 7.003e-01 21.176 < 2e-16 *** food_num 1.875e+00 5.126e-01 3.658 0.000255 *** live_num 1.269e-02 8.930e-03 1.421 0.155409 work_num -3.475e-03 5.177e-03 -0.671 0.502075 move_num 6.138e-03 9.745e-04 6.298 3.10e-10 *** AGE_001 9.442e-01 1.272e-01 7.423 1.21e-13 *** AGE_002 1.802e+00 2.266e-02 79.333 < 2e-16 *** AGE_003 1.735e+00 7.755e-02 22.366 < 2e-16 *** AGE_004 2.670e+00 9.561e-02 27.929 < 2e-16 *** AGE_005 2.590e+00 1.615e-01 16.037 < 2e-16 *** AGE_006 3.735e+00 2.749e-01 13.585 < 2e-16 *** AGE_007 -9.462e-02 5.665e-01 -0.167 0.867340 AGE_008 4.717e+00 3.769e-01 12.517 < 2e-16 *** in_hanriverpark 3.740e-02 1.687e+01 0.002 0.998231 seasonspring 5.604e+02 1.093e+01 51.282 < 2e-16 *** seasonsummer -1.745e+02 9.936e+00 -17.565 < 2e-16 *** seasonwinter -2.079e+01 1.247e+01 -1.668 0.095384 park_YN -2.664e+01 7.984e+00 -3.337 0.000849 *** school_YN1 3.392e+00 8.681e+00 0.391 0.695953 traffic 1.487e+00 2.608e+00 0.570 0.568575</pre>					<pre>Residual standard error: 393.6 on 14443 degrees of freedom Multiple R-squared: 0.9246, Adjusted R-squared: 0.9244 F-statistic: 4787 on 37 and 14443 DF, p-value: < 2.2e-16</pre>				

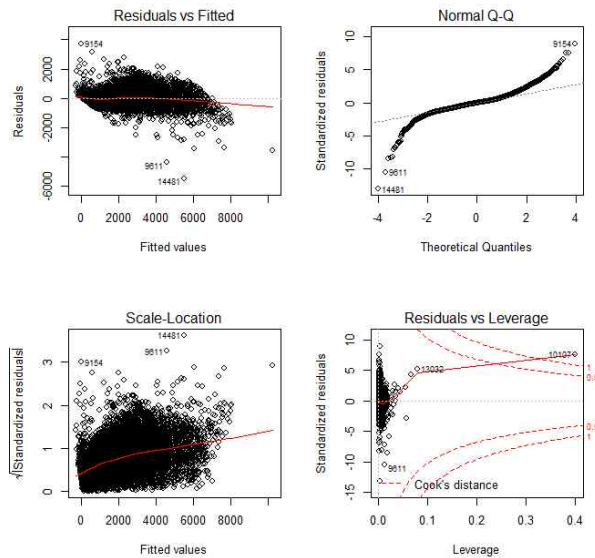
[그림 19] 위 모델에서 다중공선성 높은 'total_store'변수 제거한 모델 결과 창

vif가 높은 'total_store'를 제거하고 다시 다중공선성을 확인한 결과 food_num, AGE_003, number_of_holders:food_num, group:AGE_004, group:AGE_005의 vif값이 10 이상인 것을 확인할 수 있었다. 이 중 설명하기 어려운 number_of_holders:food_num를 제거하고 교호작용이 있다고 판단한 group:AGE_004, group:AGE_005를 선택하였다. 변수 제거 과정에서 main effect보다는 interaction effect를 제거하는 것이 더 낫기 때문이다. 하지만 제거 후에도 여전히 다중공선성이 높다고 나오는 AGE_003 변수는 main effect라도 제거했다. 그 결과 vif>10 인 변수들은 발견되지 않았다.

따라서 최종 모델의 회귀식은 다음과 같다.

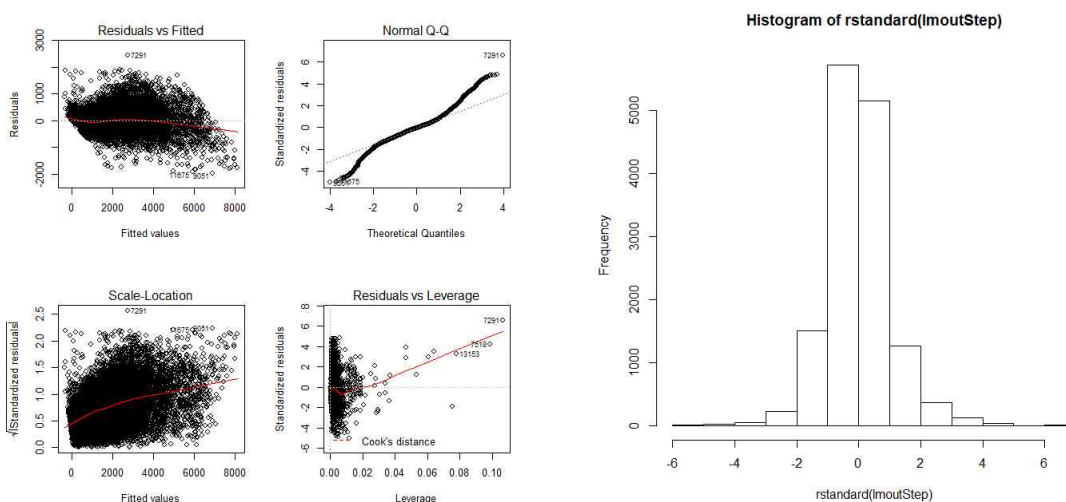
*“user_num ~ area5 + number_of_holders + group + food_num + live_num + work_num + move_num + AGE_001 + AGE_002 + AGE_004 + AGE_005 + AGE_006 + AGE_007 + AGE_008 + in_hanriverpark + season + park_YN + school_YN + traffic + number_of_holders * work_num + group * AGE_008 + group * traffic + food_num * move_num + live_num * move_num”*

이 회귀식을 가지는 모델을 stepwise로 변수선택을 진행하였다. 이 모델을 lmfinal이라고 칭하겠다.



[그림 20] 최종 모델의 회귀식을 가지는 model

이 모델의 정규성을 검증하기 Kolmogorov-Smirnov test를 진행했다. 귀무가설은 “정규분포를 따른다.”로 하고 검정한 결과 p-value가 $2.2e-16$ 보다 작아 귀무가설이 기각되었다. 따라서 정규성이 보증되지 않는다고 결론을 내렸다. 정규성을 해결하는 방법은 1) 이상치 영향치 제거 2) 종속변수 수학적 transformation 3) 새로운 독립변수 추가가 있는데, 1), 2) 방법 사용했다. 정규성을 해결하기 위하여 car 패키지의 outlierTest 함수를 이용해 n.max를 최종 회귀모델에 들어가는 데이터의 행 개수로 설정하였고, 그 결과 이상치는 38개가 나왔다. 이 이상치를 제거하고 다시 최종회귀모델식을 이용해 회귀모델(lmout)을 생성했다. 그리고 Kolmogorov-Smirnov test를 다시 진행해봤지만 여전히 귀무가설이 기각된다. 따라서 standardized 잔차를 히스토그램을 그려 정규성을 확인해봤다.

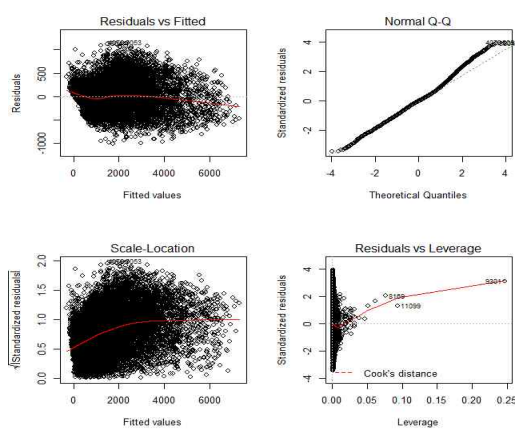


[그림 21] 이상치 제거 모델 plot [그림 22] 이상치제거모델의 표준화잔차히스토그램

표준화잔차 히스토그램은 정규분포모양과 비슷하게는 보이지만 정규분포를 따른다고 확신을 내릴 수는 없겠다. 따라서 이 모델의 종속변수(user_num)을 log변환, $\sqrt{\quad}$ 변환을 사용하고 다시 모델을 생성하여 결과를 확인해봤지만 정규성, 등분산성이 더 안 좋아졌다. 따라서 transformation 방법은 사용하지 않고, Cook's distance를 이용해 영향치를 추가적으로 제거하

기로 했다. cooks.distance 함수의 결과가 $4/n$ (n 은 최종 데이터의 행개수) 보다 크면 제거하기로 했다. 영향치를 제거한 데이터를 넣고 최종회귀식을 이용해 다시 회귀모델을 생성하고 stepwise한 최종모델을 확인한 결과 vif가 모두 6이하로 다중공선성이 높은 변수도 없고, 모델의 plot과 표준화잔차 히스토그램을 확인한 결과 정규성과 등분산성을 만족한다는 것을 확인할 수 있었다. 따라서 이 모델(lmfStep)을 최종회귀모델로 결정하였다.

최종회귀모델의 결과는 다음과 같다. [그림], [그림], [그림]을 통해 최종회귀모델이 정규성과 등분산성을 만족한다는 것을 확인할 수 있고, [그림]을 통해 독립성도 만족한다는 것을 확인할 수 있다.



```
Call:
lm(formula = user_num ~ area5 + number_of_holders + group + food_num +
    move_num + AGE_002 + AGE_004 + AGE_005 + AGE_006 + AGE_008 +
    in_hanriverpark + season + traffic + group:AGE_008 + group:traffic +
    food_num:move_num, data = df4)

Residuals:
    Min       1Q   Median       3Q      Max
-1010.75  -190.31    -6.04   161.05  1159.47

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.848e+01  1.411e+01  -6.270 3.72e-10 ***
area5동남권   1.534e+01  9.721e+00   1.578 0.114681
area5동북권   4.872e+01  9.739e+00   5.003 5.72e-07 ***
area5서남권   6.301e+01  9.534e+00   6.609 4.02e-11 ***
area5서북권  -3.931e+01  1.089e+01  -3.611 0.000306 ***
number_of_holders -5.458e+00  6.130e-01  -8.904 < 2e-16 ***
group         9.279e+00  5.476e-01  16.945 < 2e-16 ***
food_num      1.747e+00  1.822e-01   9.591 < 2e-16 ***
move_num      5.366e-03  6.655e-04  8.064 8.02e-16 ***
AGE_002       2.378e+00  1.538e-02  152.600 < 2e-16 ***
AGE_004       3.655e+00  5.860e-02  62.363 < 2e-16 ***
AGE_005       2.712e+00  1.025e-01  26.458 < 2e-16 ***
AGE_006       3.266e+00  2.219e-01  14.723 < 2e-16 ***
AGE_008       9.005e+00  3.550e-01  25.368 < 2e-16 ***
in_hanriverpark1 2.386e+01  1.422e+01   1.678 0.093457 .
seasonspring   5.299e+02  8.445e+00  62.745 < 2e-16 ***
seasonsummer  -1.335e+02  7.759e+00 -17.203 < 2e-16 ***
seasonwinter   2.560e+01  9.643e+00   2.655 0.007937 **
traffic        -2.611e+00  2.062e+00  -1.266 0.205477
group:AGE_008  -1.256e-01  6.451e-03 -19.473 < 2e-16 ***
group:traffic  -6.738e-01  1.415e-01  -4.761 1.95e-06 ***
food_num:move_num -8.449e-05  1.338e-05  -6.315 2.79e-10 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 293.4 on 13393 degrees of freedom
Multiple R-squared:  0.9475,    Adjusted R-squared:  0.9474
F-statistic: 1.151e+04 on 21 and 13393 DF,  p-value: < 2.2e-16
```

[그림 27] 최종회귀모델 결과창

최종회귀모델의 모형이 적합한지 확인하기 위해 F-test를 시행하였다.

‘귀무가설 : 이 모형은 적합하지 않다. vs 대립가설 : 이 모형은 적합하다.’로 가설을 설정하였다. 유의수준 0.05하에 F검정을 시행하니 p-value가 $2.2e-16 < 0.05$ 이므로 귀무가설을 기각하여 이 모형이 적합함을 알 수 있었다.

최종회귀모델의 모형이 적합하다고 판단되었으므로 각 독립변수의 유의성을 확인하기 위해 t test를 시행하였다.

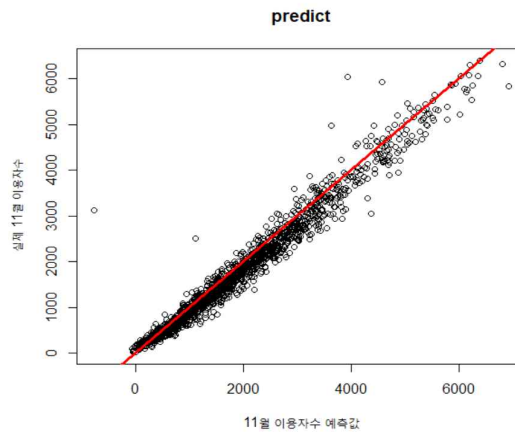
‘귀무가설 : 회귀계수 $B_i=0$ (독립변수 X_i 는 유의하지 않다.) vs 대립가설 : 회귀계수 $B_i \neq 0$ (독립변수 X_i 는 유의하다.)’로 가설을 설정하고 유의수준 0.05하에서 t-test를 시행하였다. 그 결과, area5\$동남권, traffic(교통변수)의 p-value가 0.05보다 커서 귀무가설을 기각하지 못했다. 따라서 area5\$동남권과 교통은 이용자수에 크게 영향을 주지 않음을 알 수 있었다.

2. 회귀모델 데이터 예측

앞서, 2019년 1월 ~ 2019년 10월의 데이터로 최종 회귀모델 (lmfStep)을 통해 예측을 하기 위해 2019년 11월 데이터를 예측 데이터로 생성하였다. predict()함수를 통해 11월 데이터와 회귀 최종모델을 넣고 예측을 시행하였다.

plot과 신뢰구간, 2가지 방법으로 예측이 잘 되었는지 확인해보았다.

1). plot을 활용해 최종회귀모델의 예측력 확인하기



[그림 28] \hat{y} vs y plot

x축에는 11월 이용자 수를 예측한 값을, y축에는 11월 이용자 수의 실제값을 넣고 scatter plot을 그려보았다. 이는 $y=x$ 축과 거의 유사한 모양으로 그려졌으므로 11월 이용자수의 예측 값이 실제 11월 이용자 수에 거의 가깝게 예측되었음을 알 수 있었다.

2) 신뢰구간을 활용해 예측의 성공률 확인하기

plot을 통해서는 예측의 성공여부를 시각적으로 확인할 수 있었다. 이를 정확한 수치 값으로 확인하기 위해 신뢰구간을 활용하였다. 11월 이용자 수가 예측된 값이 [신뢰구간의 하한, 신뢰구간의 상한] 안에 포함된 개수를 세어 sum에 저장하였다. 그리고 11월 이용자수의 실제값을 real 변수에 저장하였다.

그리고 예측의 성공여부를 비율로 확인하니 $(\text{sum}/\text{real}) = 0.965$ 의 값으로 96.5%로 예측에 성공했음을 알 수 있었다.

3) 10-folds cross validation을 통한 회귀모델 성능 확인

최종회귀모델(lmfStep)의 성능을 평가하기 위해 k-fold cross validation(교차검증)을 시행하였다. k=10개의 fold를 만든 후, 교차검증을 시행하여 각 fold별로 adjusted R square, aic, bic 값을 확인 해보았다.

	adjR2	aic	bic
Fold01	0.9408529	17979.658	18062.889
Fold02	0.9474775	13487.300	13570.519
Fold03	0.9469216	16731.318	16814.524
Fold04	0.9495493	4675.407	4758.638
Fold05	0.9499959	15013.615	15096.846
Fold06	0.9421382	14503.709	14586.939
Fold07	0.9496038	14466.160	14549.367
Fold08	0.9461416	10238.409	10321.628
Fold09	0.9465470	12962.949	13046.179
Fold10	0.9474201	16801.656	16884.899

[그림 29] 10-folds cross validation 결과

그리고 10개의 fold의 adjusted R square, aic, bic의 평균값과 adjusted R square, aic, bic의 분산값을 확인하였다. 이를 통해 모델의 평균 성능과 모델의 성능 분산을 알 수 있었다.

	[,1]	[,2]	[,3]
mean	0.9466648	13686.02	13769.24
sd	0.003053357	3859.988	3859.988

[그림 30] 모델의 성능 분산(오차)

이는 차례로 adjusted R square, aic, bic의 평균 성능 (1행), adjusted R square, aic, bic의 성능 분산(2행)이다. 최종모델을 교차검증한 결과, adjusted R square값이 0.947, aic가 13686,

bic가 13769가 나왔음을 알 수 있다.

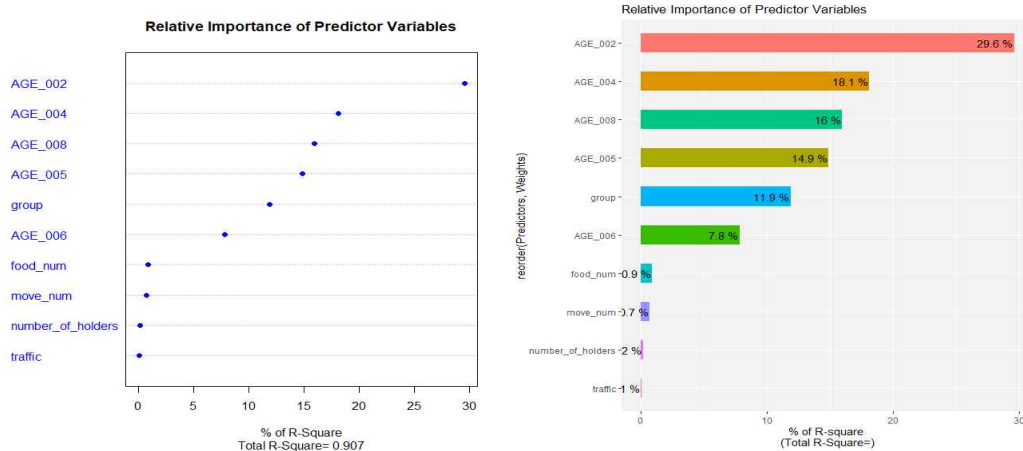
제3장 결론

제1절 분석결과 정리

최종회귀모델의 식은 다음과 같다.

$$\begin{aligned} \hat{Y} = & (-8.848e+01) + (1.534e+01)area5_남권 + (4.872e+01)area5_동북권 + (6.301e+01)area5_서남권 \\ & + (-3.931e+01)area5_서북권 + (-5.458e+00)number_of_holders + (9.279e+00)group \\ & + (1.747e+00)foodnum + (5.366e-03)movenum + (2.378e+00)AGE002 + (3.655e+00)AGE004 \\ & + (2.712e+00)AGE005 + (3.266e+00)AGE006 + (9.005e+00)AGE008 + (2.386e+01)inhanriverpark1 \\ & + (5.299e+02)seasonspring + (-1.335e+02)seasonsummer + (2.560e+01)seasonwinter \\ & + (-2.611e+00)traffic + (-1.256e-01)group:AGE008 + (-6.738e-01)group:traffic \\ & + (-8.449e-05)foodnum:movenum \end{aligned}$$

그리고 최종회귀 모델의 평균 성능은 adjR^2 기준으로 0.9467로 유의하게 나왔다. 또한 위에서 확인했듯이 정규성, 등분산성, 독립성도 만족한다. 이 회귀 모델의 변수 중요도를 확인해보자. 단, 변수중요도는 수치형변수만 가능하다는 단점이 있다.



[그림 31, 32] 최종회귀모델의 변수 중요도

변수중요도를 확인한 결과 대체로 지난 월의 연령대별 이용자수가 영향력이 있고, 그 중 20대 이용자수가 가장 큰 영향을 주는 것을 확인했다. 또한 단체권, 대여소 주변 음식점 수, 유동인구 수, 거치대 수, 교통수단 개수(지하철 역, 버스정류장 수)도 어느 정도 중요한 영향을 끼치는 것으로 판단되었다. 따라서 우리는 추정된 회귀식과 변수중요도를 토대로, 따릉이 이용자수는 20대 이용자의 영향을 많이 받고, 지역에 따라서는 별 차이를 보이지는 않지만, 다른 변수의 값이 같을 때, 서남권이 다른 지역보다 이용자수가 많은 영향을 끼친다고 말할 수 있겠다. 우리는 이 모델을 토대로, 주변 환경의 특성과 계절, 지역, 전월의 연령별 이용자수, 교통시설수, 한강공원유무 등을 토대로 이용자수를 예측해 시민의견수렴과 업체 측 의견 반영 사이의 기간을 단축해 보다 원활한 운영을 기대한다.

1) 최적의 변수개수 확인 플랏

