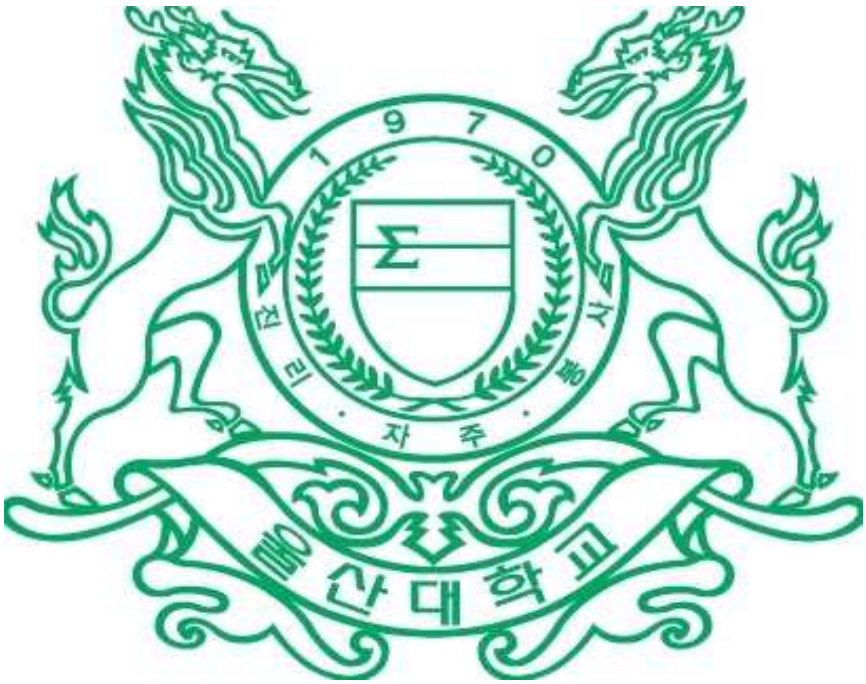
Term ProJect

- 데이터 사이언스 텀 프로젝트 -



학과	IT융합학과
분반	[01분반]
과목명	데이터 사이언스
학번/이름	20184256 김연수 20184257 김연주 20184258 박선재
담당 교수	권영근 교수님
제출 일	2021년 06월 11일

울산대학교
UNIVERSITY OF ULSAN

<목 차>

1. 도입

1.1 도입

서울시는 공공 운영 서비스로 공공자전거 '따릉이'를 운영중에 있다. 현재 약 20,000대의 '따릉이' 자전거를 보유하고 있으며, 2017년 기준 하루 평균 약 1만 4천건이 운행되고 있다. 이용자의 절반이 20대이며 20대 서비스 사용자 94.8%가 '따릉이' 회원가입을 하였을 정도로 서울시에서 가장 활발하게 운영되고 있는 서비스 중 하나다. 우리는 이러한 가입자별 연령대, 성별, 일자별 공공자전거 신규가입자 추이를 확인함으로써 자전거 대여사업 운영자 입장에서 관리를 용이 하게 할 수 있을 것이다.

[예 시] 특정 일자에 가입자수의 증가추이를 통해서 이용객을 어느 정도 예측할 수 있고, 이용자수를 예측 및 파악함으로써 사전에 해당 위치에 자전거 등을 배치 할 수 있다.

2. 가설

2.1 가설

* 가설 1 : 공공자전거 가입자수는 남자가 여자보다 많을 것이다.

* 가설 2 : 나이 대별 가입자수는 20대가 최대, 30대 -> 10, 40 ,50 ,60 ,70 순서로 감소할 것이다.

외국인의 경우 성별을 확인 할수 없다. 국내인의 비율상으로봤을때는 맞게 예측했다.

3. 데이터 설명

3.1 데이터

서울특별시 공공자전거 신규가입자 정보(일별, 2020 9월 ~ 2021 1월)

데이터 출처 : http://data.seoul.go.kr/dataList/OA-15243/S/1/datasetView.do

Saving 20184256-data1.csv to 20184256-data1 (2).csv

Saving 20184256-data2.csv to 20184256-data2 (2).csv

Saving 20184256-data3.csv to 20184256-data3 (2).csv

Saving 20184256-data4.csv to 20184256-data4 (2).csv

Saving 20184256-data5.csv to 20184256-data5 (2).csv

-> 년도 별 가입자별 , 사용자코드 , 연령대코드, 성별, 가입 수

3.2 데이터 의미

- 가입일자 : 실제 가입일자.
- 사용자 코드 : 유저의 특성을 나타낸 코드.
- 연령대 코드 : 신규가입자의 연령을 나타냄 해당 값은 아래의 인덱스를 참조.
- 성별 : F,M (여자,남자)
- 가입수 : 실제 가입수

가입일자	사용자코드	연령대코드	성별	가입 수
2020-09-01	USR_001	AGE_001	F	179
2020-09-01	USR_001	AGE_001	M	180
2020-09-01	USR_001	AGE_002	F	564
2020-09-01	USR_001	AGE_002	M	610
2020-09-01	USR_001	AGE_003	F	219
2020-09-01	USR_001	AGE_003	M	246
2020-09-01	USR_001	AGE_004	F	147
2020-09-01	USR_001	AGE_004	M	153
2020-09-01	USR_001	AGE_005	F	102
2020-09-01	USR_001	AGE_005	M	91

... (2290 rows omitted)

- AGE_001 : 10대
- AGE_002 : 20대
- AGE_003 : 30대
- AGE_004 : 40대
- AGE_005 : 50대
- AGE_006 : 60대
- AGE_007 : 70대
- AGE_008 : 비회원, 외국인, 그 이상 등

연령대에 해당하는 값을 변환하여 추가

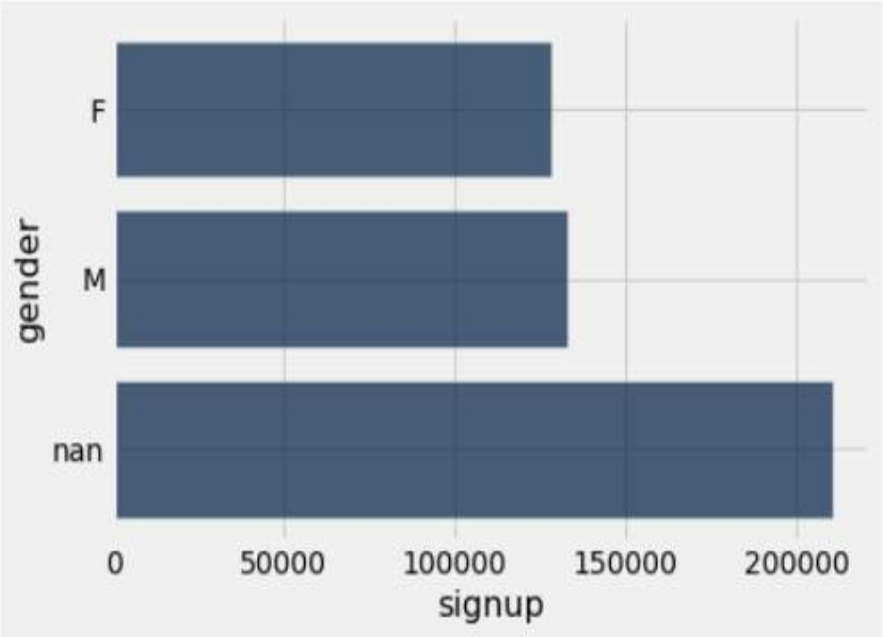
가입일자	사용자코드	연령대코드	성별	가입 수	나이
2020-09-01	USR_001	AGE_001	F	179	10s
2020-09-01	USR_001	AGE_001	M	180	10s
2020-09-01	USR_001	AGE_002	F	564	20s
2020-09-01	USR_001	AGE_002	M	610	20s
2020-09-01	USR_001	AGE_003	F	219	30s
2020-09-01	USR_001	AGE_003	M	246	30s
2020-09-01	USR_001	AGE_004	F	147	40s
2020-09-01	USR_001	AGE_004	M	153	40s
2020-09-01	USR_001	AGE_005	F	102	50s
2020-09-01	USR_001	AGE_005	M	91	50s

... (2290 rows omitted)

4. 분석 방법 및 결과

4.1 분석 방법 및 결과

* 가설 1. 가입자 수는 남자가 여자보다 많을 것이다.



gender	signup
F	128366
M	132627
nan	210378

남성과 여성의 성별보다 nan값이 더 많이 찍히는 것을 볼 수 있는데, 어떤 데이터 때문인지 살펴보기 위해서 데이터 핸들링을 통해서 판다스 데이터 프레임으로 변환하였다.

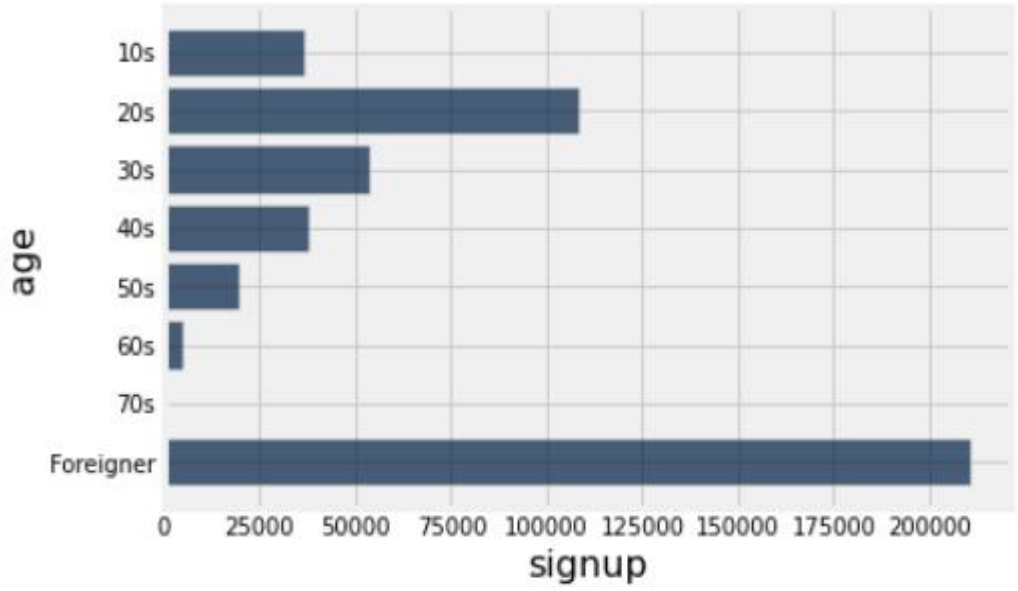
	가입일자	사용자코드	연령대코드	성별	가입 수	나이
0	2020-09-01	USR_001	AGE_001	F	179	10s
1	2020-09-01	USR_001	AGE_001	M	180	10s
2	2020-09-01	USR_001	AGE_002	F	564	20s
3	2020-09-01	USR_001	AGE_002	M	610	20s
4	2020-09-01	USR_001	AGE_003	F	219	30s
...	...	...	...	...	...	...
2295	2021-01-31	USR_001	AGE_006	F	13	60s
2296	2021-01-31	USR_001	AGE_006	M	17	60s
2297	2021-01-31	USR_001	AGE_007	F	1	70s
2298	2021-01-31	USR_001	AGE_007	M	3	70s
2299	2021-01-31	USR_001	AGE_008	M	1	Foreigner

2300 rows × 6 columns

nan값이 찍히는 연령대를 살펴보니 모두 외국인 데이터이다. 국내인의 비율상으로 봤을때

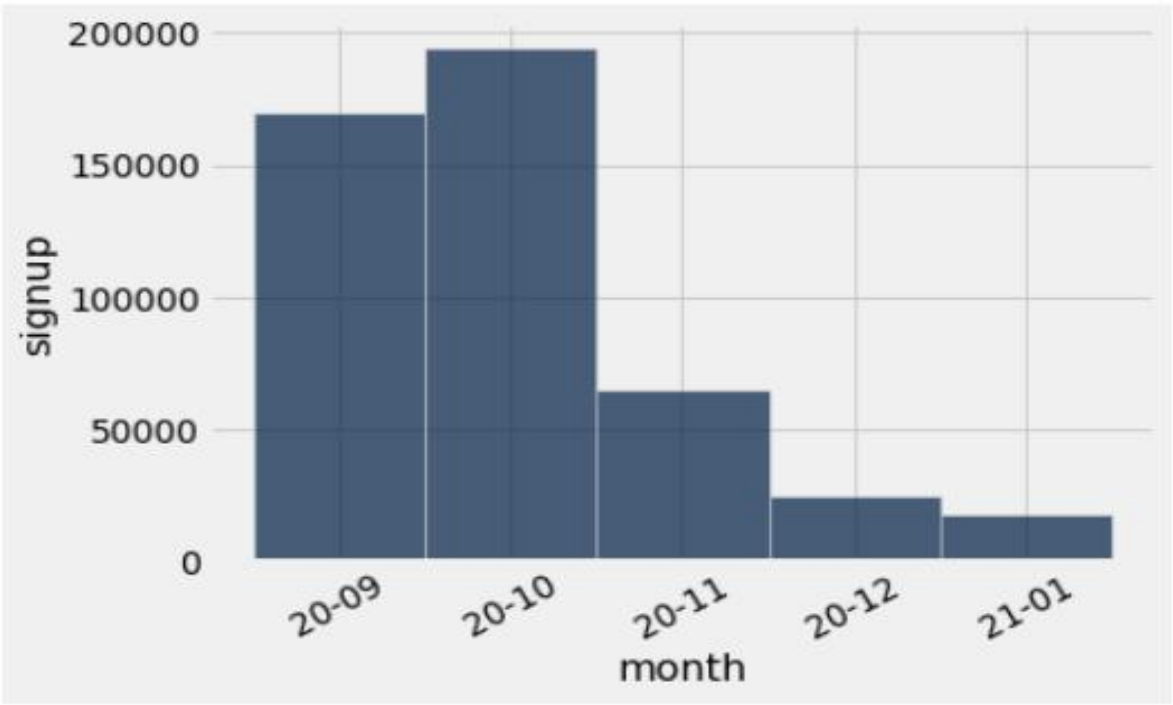
는 맞게 에측 했지만 큰 차이는 없는 것으로 보인다. 국내에 거주중인 외국인의 경우는 성별을 확인 할 수 없기 때문에 가설에 대입할 수 없다.

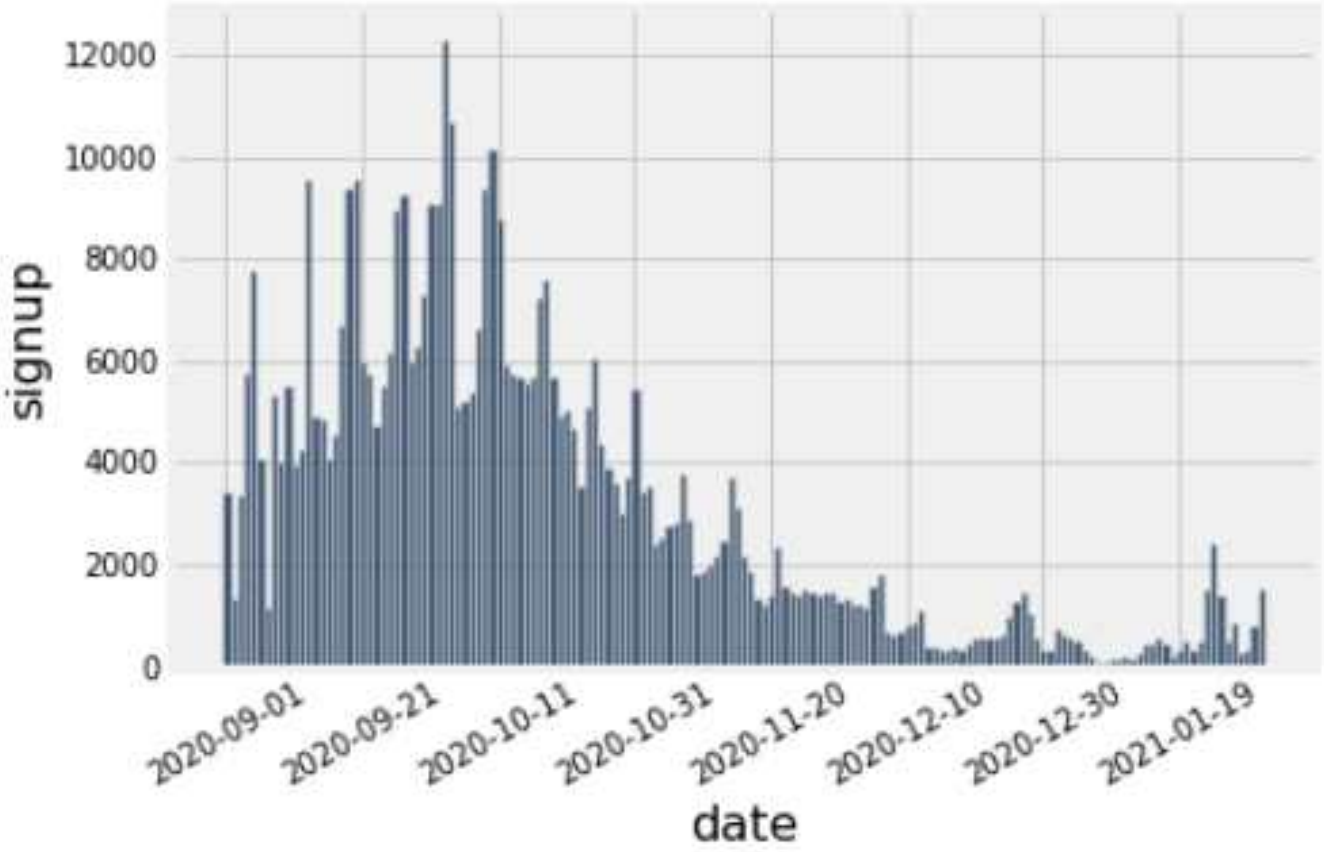
* 가설 2. 나이대 별로 가입자수가 다를 것이다.
(20대가 최대, 30→10→40 ,50 ,60 ,70 순서로 감소 할 것이다.)



20대의 가입자 수가 가장 많았으며, 30대>40대>10대>50대>60대 순서로 나타났다.
* 추가 가설. 계절에 따른 가입자 수도 다를 것이며, 겨울에 가입자 수가 더 낮을 것이다.

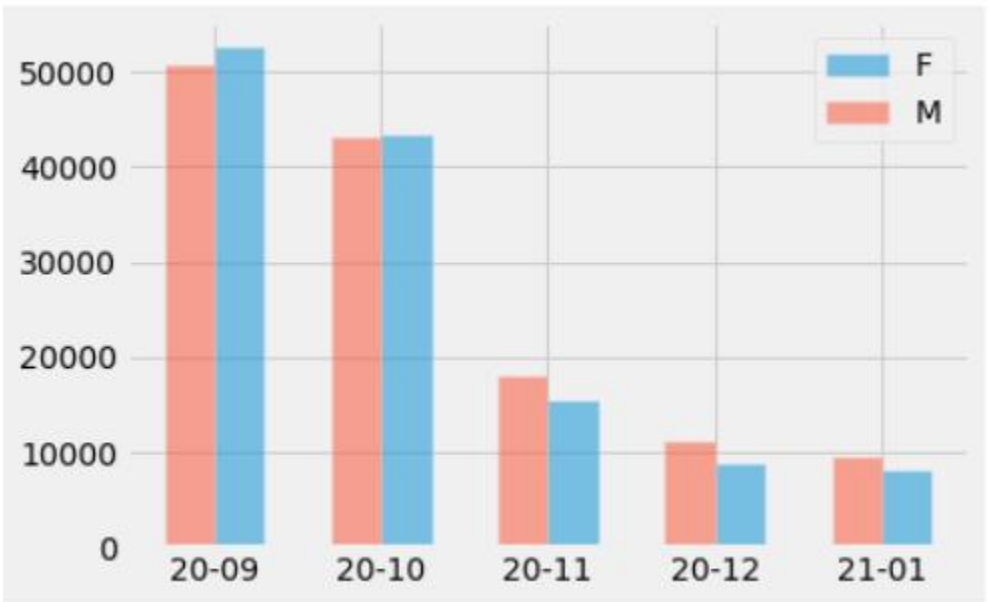
월별 가입자 수를 확인하기 위해 판다스 데이터 프레임에서 전처리를 해준 후 다시 table 형식으로 바꿔준다.





가입자 수의 분포를 보면 계절의 영향을 크게 받는다. 겨울철의 경우 급격하게 감소하는 경향을 보인다.

남녀의 월별 가입수



겨울철의 경우 남성의 가입자가 더 많고, 가을철의 경우 여성가입자가 더 많다.

5. 결과

5.1 결론

2020년 9월에서 2021년 1월까지 서울특별시 공공자전거 따릉이의 신규가입자 수의 성별, 날짜, 나이대 별로 예측해 보았다. 계절에 따라 공공자전거의 사용에 계절과 성별이 영향을 미칠 수 있다고 판단되었다. 분석한 결과 겨울철의 경우 남성의 가입자가 수가 더 많고, 가을철의 경우 여성 가입자의 수가 더 많았다. 계절과 성별에 따른 이용량이 크게 다르지 않을 것 같았지만 생각과 다른 결과가 나왔다.

5.2 논의사항 및 개선사항

날씨, 공휴일 등의 변수가 있음을 확인 했으며, 실제로 연관지어 분석한다면 더욱 좋은 예측 성능을 보일 것이다. 가입자수의 분포를 보면 계절의 영향을 크게 받는다고 생각한다. 그래서 겨울철의 경우 급격하게 감소하는 경향을 보인다. 겨울철 공공자전거의 사용량이 급격하게 줄어든 것을 확인할 수 있었다. 그에 따라 서울내의 영역별 자전거의 배치수를 줄이거나, 비용감소를 통해 사용량을 늘리는 방안을 고려해 볼 수 있을 것이다. 위치별 사용량을 확인 할 수 없었기 때문에 겨울철에 서울 영역별 사용량을 측정하기 어려운 점이 있었다. 이것을 알 수 있었다면 겨울철에 사용량이 적은 영역에는 배치수를 줄일 때 도움이 될 수 있었을 것이다. 그리고 사용량이 적지만 그 중 자전거를 사용한 이용자가 많은 지역을 알 수 있다면 그 지역에 자전거를 더 배치 하는 방안도 생각 해 볼 수 있었을 것이다.