

서울시 자치구별 인구수와

자치구별 지하철 이용수의 상관관계

20172287 윤소영

A반 | 데이터베이스 설계 및 보안 | R 데이터 분석

1. 주제 선정

**1. 데이터 선정 이유**

평소 대중교통을 많이 이용하고, 많은 사람들도 대중교통을 이용하고 있기 때문에 친근한 데이터를 선정하게 되었습니다.

**2. 데이터 출처 – 서울 열린데이터 광장**

a. 서울시 자치구단위 서울 생활인구 일별 집계표

<http://data.seoul.go.kr/dataList/datasetView.do?infId=OA-15379&srvType=S&serviceKind=1&currentPageNo=1>

b. 서울시 지하철호선별 역별 승하차 인원 정보

<http://data.seoul.go.kr/dataList/datasetView.do?infId=OA-12914&srvType=S&serviceKind=1&currentPageNo=1>

c. 서울교통공사 지하철역 주소 및 전화번호 정보

<http://data.seoul.go.kr/dataList/datasetView.do?infId=OA-12035&srvType=S&serviceKind=1&currentPageNo=1>

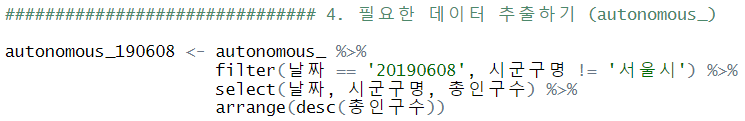
**3. 가설 설정**

인구수가 많은 자치구에 있는 역이 많이 이용될 것이다.

2. 자치구별 인구수 분석

**1. 자치별 인구수**

a. 원본 데이터에서 필요한 데이터만 추출하기



엑셀 파일을 읽고, CSV파일로 만들어준 후 rename으로 데이터의 열 제목을 변경해준다. 가장 최근인 2019년 06월 08일의 자치구별 총 인구수를 내림차순으로 정렬하여 추출한다. 그 과정에서 서울시 전체 인구수는 제외한다.

b. 추출 후 행과 열의 변화

스크린샷이(가) 표시된 사진

매우 높은 신뢰도로 생성된 설명

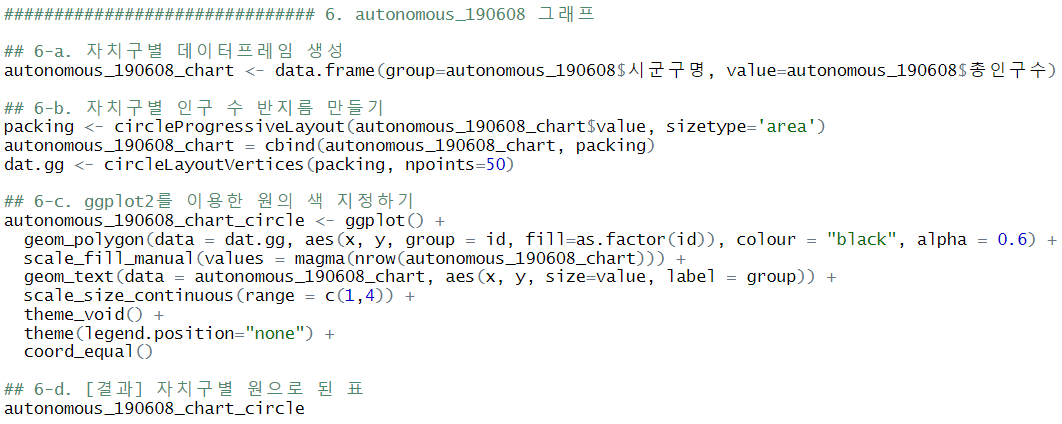
c. 추출 데이터



**2. 자치별 인구수 그래프**

a. 추출 데이터 그래프

2-1의 데이터를 이용해 자치구별 인구수 그래프를 만든다.



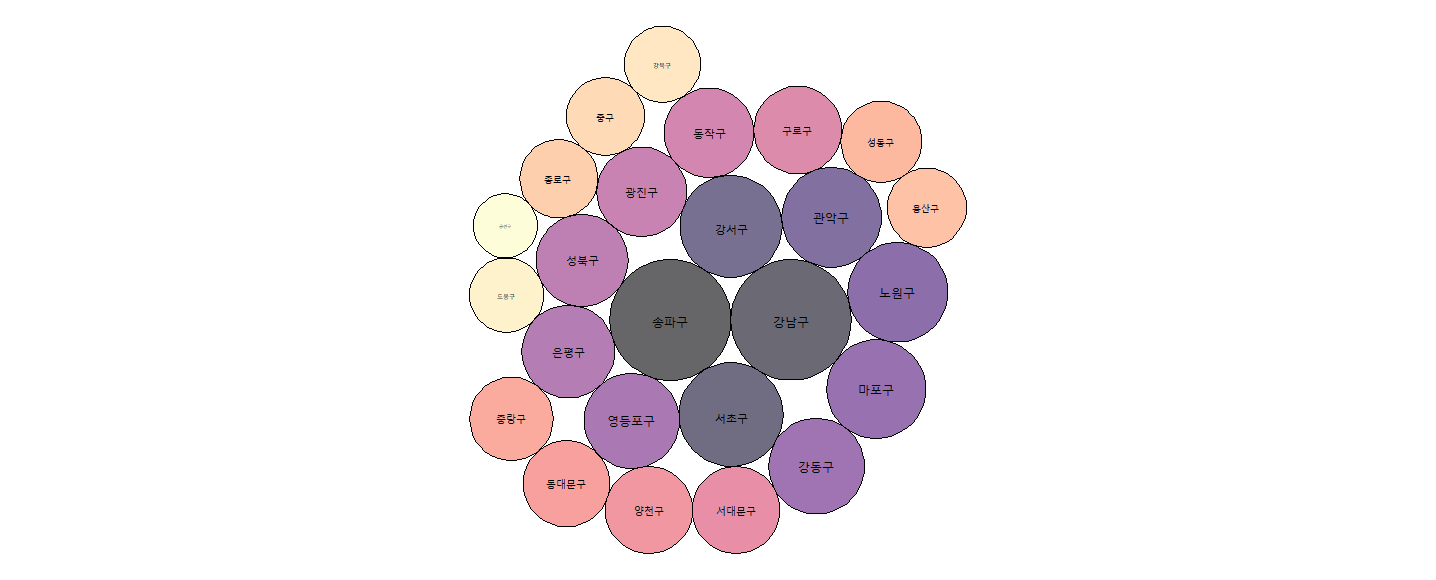
**[그래프 코드1] 2-1에서 추출한 데이터를 이용해 원형 그래프를 만들기 위한 코드**

[그래프 코드1]에 있는 6-a를 보면 자치구별 데이터를 묶어 새로운 데이터 프레임을 만든 후 autonomous\_190608\_chart 라는 이름으로 지정한다.

다음으로 6-b를 보면 패키지 ‘packcircles’ 패키지를 이용해 ‘circleProgressiveLayout’로 원의 레이아웃을 생성하고, 테이블의 열을 합치기 위해 cbind를 사용하였다.

마지막으로 ggplot2 패키지를 이용해 원하는 색으로 그래프를 만들어낸다.

b. 결과물

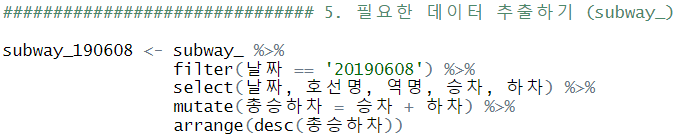


송파구 > 강남구 > 서초구 > 강서구 순으로 인구가 많다는 것을 알 수 있다. 가설을 적용했을 때 송파구에 있는 지하철역을 가장 많이 이용될 것이다.

3. 역별 이용자 수 분석

**1. 역별 이용자 수**

a. 데이터 추출



엑셀 파일을 읽고, CSV파일로 만들어준 후 rename으로 데이터의 열 제목을 변경해준다. 앞의 인구수를 2019년 06월 08일 날짜로 분석했기 때문에 마찬가지로 같은 날짜의 역별 이용자 수를 분석한다.

b. 추출 후 행과 열의 변화

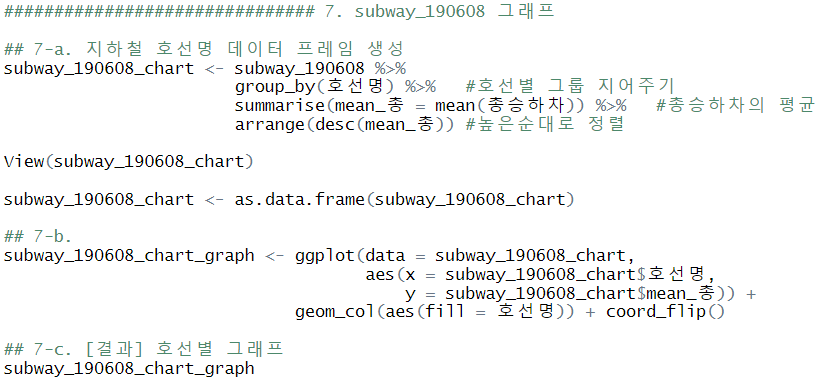


c. 추출 데이터



**2. 호선별 원형 그래프**

a. 호선별 이용수 그래프



**[그래프 코드2] 3-1에서 추출한 데이터를 이용해 원형 그래프를 만들기 위한 코드**

[그래프 코드2]에 있는 7-a를 보면 원하는 그래프를 만들기 위해 새로운 데이터 프레임을 생성하였다. 호선별 그룹으로 묶고, 총 승하차의 평균을 구하여 테이블열에 추가한 다음 높은 순대로 정렬하였다.

b. 결과물

스크린샷이(가) 표시된 사진

높은 신뢰도로 생성된 설명

1호선 > 2호선 > 4호선 > 경인선 > 3호선 순으로 이용이 많은 것을 알 수 있다. 가설을 적용하였을 때, 1호선에 있는 역 중 가장 많이 이용되는 역의 주소는 송파구일 것이다.

**3. 상위 4개역 중 각 상위 5개의 역**

a. 호선별 승차객이 많은 역 5순위

가설을 적용하기 위해 상위 호선의 4개 중 각 상위 5개의 역을 알아본다. 여기서 구한 데이터를 뒤의 4. 역별 주소 분석 에서 합칠 것이다. 하지만 경인선의 주소 데이터를 구할 수 없으므로, 경인선을 제외한 1순위인 2호선, 그 뒤로 1호선, 4호선, 3호선 4개의 역을 알아보겠다.

b. (1위) 2호선 총 승하차객이 많은 역 5순위



데이터를 추출하여 알아본다.



**[2호선 상위5개] 3-3b의 데이터 추출결과**

2호선 상위5개를 보았을 때, 2호선의 이용률이 높은 상위 5개 역은 홍대입구 > 강남 > 잠실 > 신림 > 신도림 인 것을 알 수 있다.

c. (2위) 1호선 총 승하차객이 많은 역 5순위



데이터를 추출하여 알아본다.



**[1호선 상위5개] 3-3c의 데이터 추출결과**

1호선 상위5개를 보았을 때, 1호선의 이용률이 높은 상위 5개 역은 서울역 > 종로3가 > 종각 > 종로5가 > 청량리 인 것을 알 수 있다.

d. (3위) 4호선 총 승하차객이 많은 역 5순위



데이터를 추출하여 알아본다.



**[4호선 상위5개] 3-3d의 데이터 추출결과**

4호선 상위5개를 보았을 때, 4호선의 이용률이 높은 상위 5개 역은 명동 > 혜화 > 수유 > 회현 > 미아사거리 인 것을 알 수 있다.

e. (4위) 3호선 총 승하차객이 많은 역 5순위



데이터를 추출하여 알아본다.



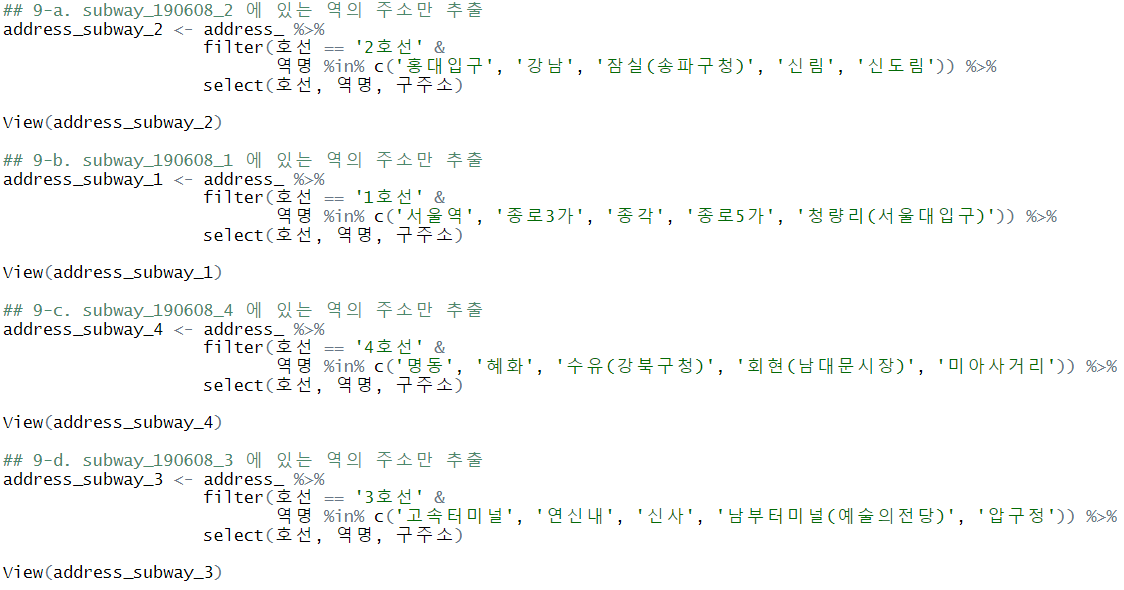
**[3호선 상위5개] 3-3e의 데이터 추출결과**

3호선 상위5개를 보았을 때, 3호선의 이용률이 높은 상위 5개 역은 고속터미널 > 연신내 > 신사 > 남부터미널 > 압구정 인 것을 알 수 있다.

4. 역별 주소 분석

**1. 역별 주소**

a. 데이터 추출



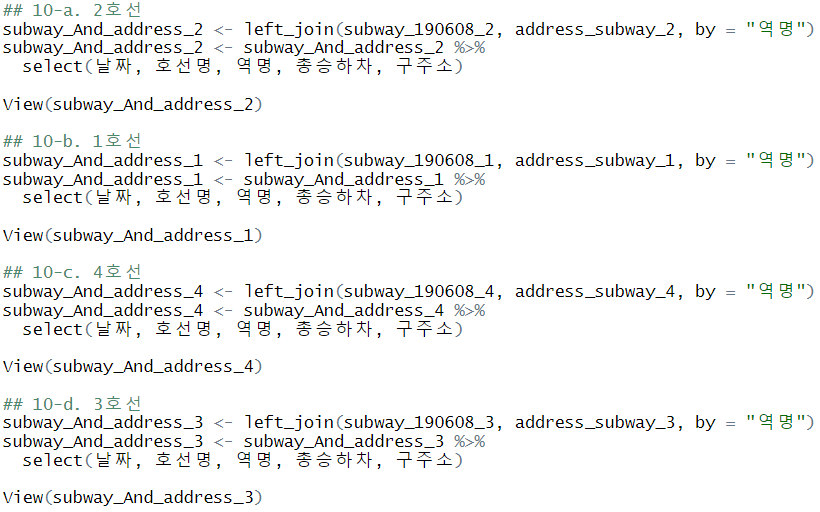
엑셀 파일을 읽고, CSV파일로 만들어준 후 rename으로 데이터의 열 제목을 변경해준다. [3.역별 이용자 수]에서 조사한 데이터에서 각 호선별 상위 5개 역의 주소 데이터를 추출한다.

b. 추출 데이터



**2. [3.역별 이용자 수] 와 4-1 데이터 조인**

a. 데이터 조인



데이터를 역명을 기준으로 오른쪽으로 조인해준다.

b. 결과

<2호선>



<1호선



<4호선>



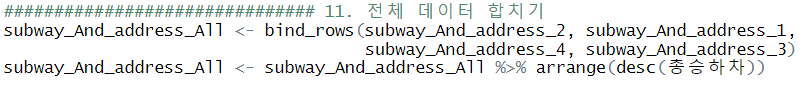
<3호선>



5. 결론

**1. 결론을 위한 데이터 합치기**

a. 데이터 추출



[4. 역별 주소 분석]의 결과를 모두 합 쳐주고, 총승하차가 높은 순위로 정렬한다.

b. 결과



**2. 결론**

호선별 많이 이용하는 순위의 경인선을 제외한 1위 ~ 4위의 1호선, 2호선, 3호선, 4호선 데이터 중 자치구별 인구 순위의 1위 ~ 4위까지의 **송파구** 지하철역 중 **1개**, **강남구** 지하철역 중 **3개**, **서초구** 지하철역 중 **2개**, **강서구** 지하철역 중 **0개** 총 6개 만이 존재하는 것을 보았을 때 **생활인구수가 높은 자치구에 위치한 역 이용수는 관련이 없다.**