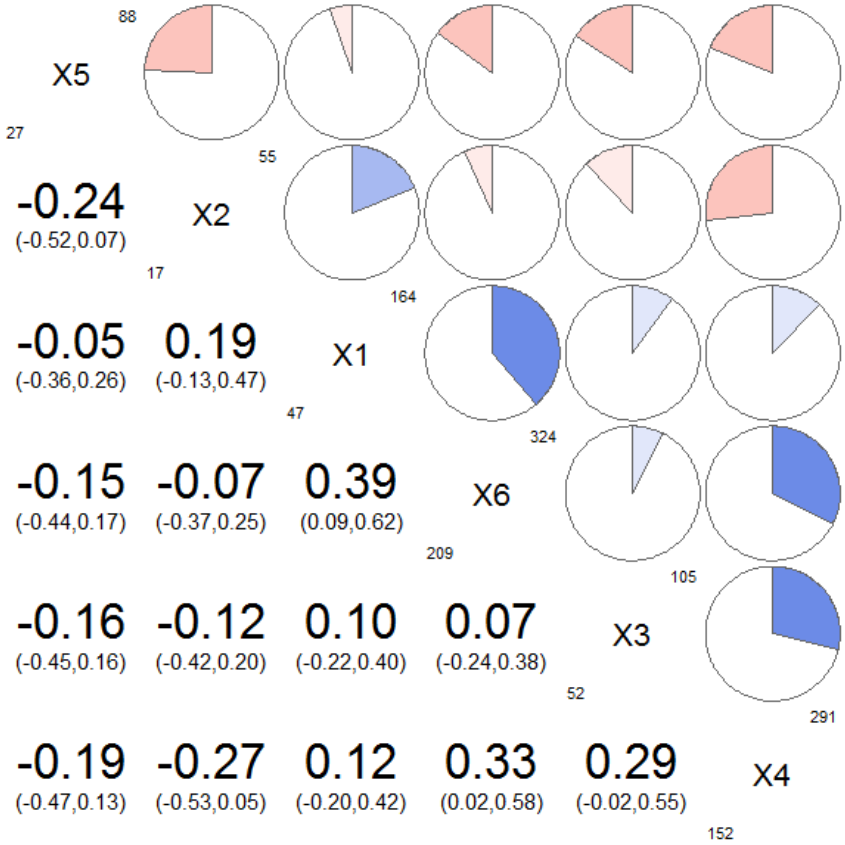
데이터 시각화

예제 1

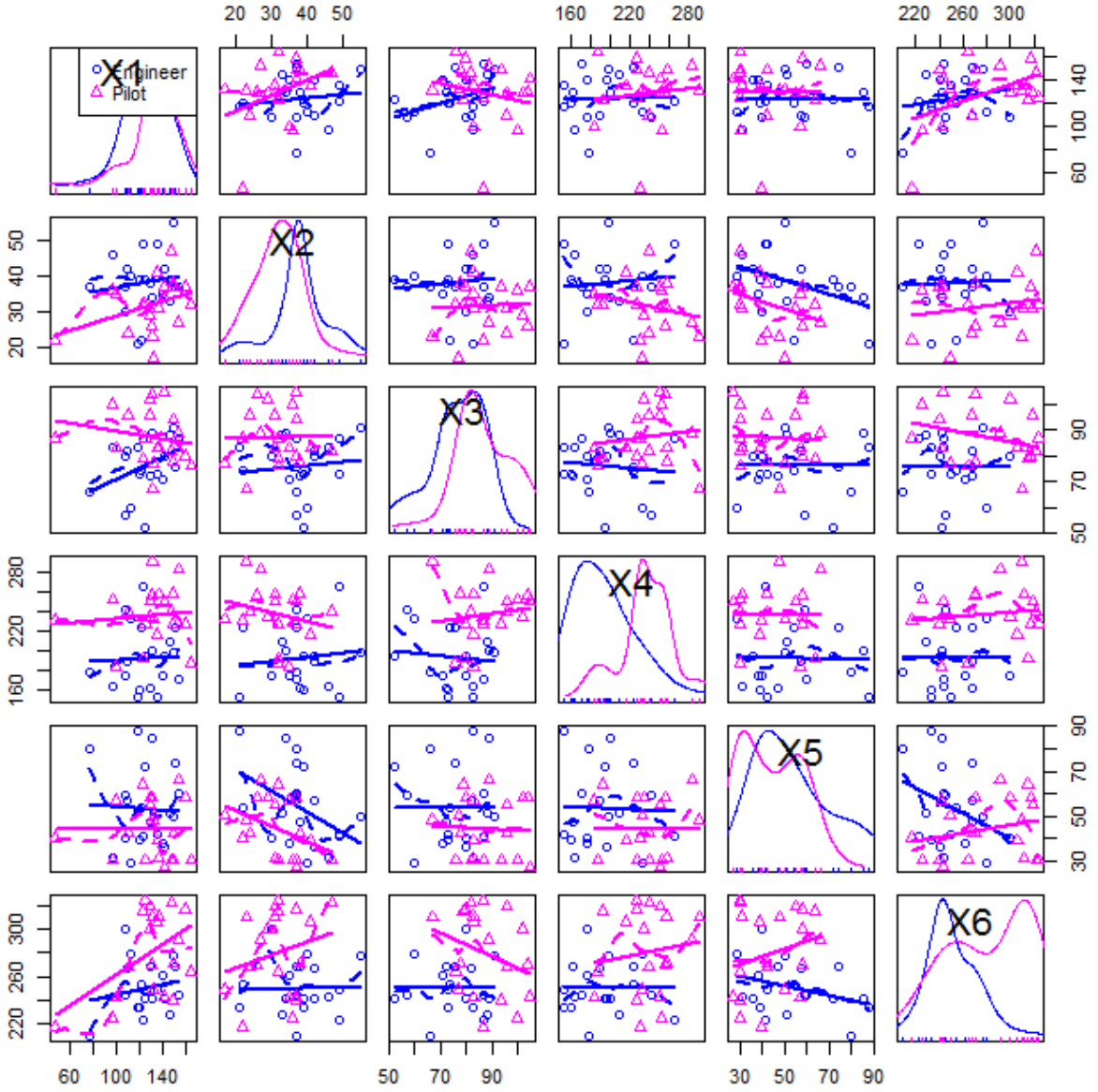
- "pilot.csv"는 엔지니어 견습생 20명과 비행조종사 20명의 여섯 가지의 테스트 결과이다
  - G : 1-Engineer, 2-Pilot
  - X_1 : 지능(intelligence)
  - X_2 : 상황설명능력(form relations)
  - X_3 : 동력측정검력계(dynamometer)
  - X_4 : 상세 표시 능력(dotting)
  - X_5 : 지각 기구 좌표와 능력(sensory motor coordination)
  - X_6 : 인내력(perseveration)

– X1~X6간의 관계를 살펴보기 위한 상관도를 작성하시오.

- 대각성분 : 변수이름과 최대값, 최소값
- 대각선 위 : 상관계수 원도표
- 대각선 아래 : 상관계수와 신뢰구간



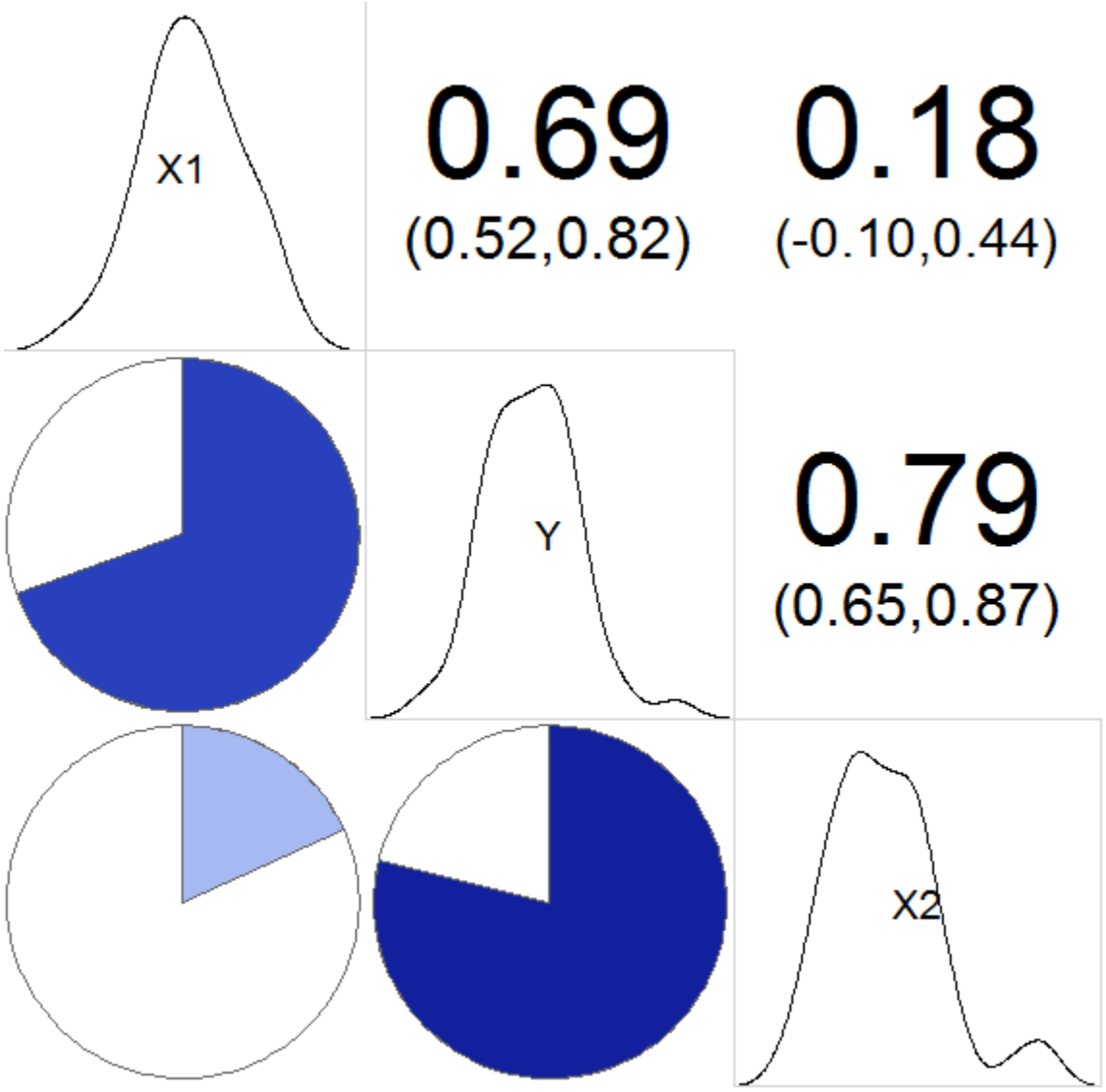
– 집단별 X1~X6 간의 관계를 살펴보기 위한 산점도 행렬을 작성하시오.



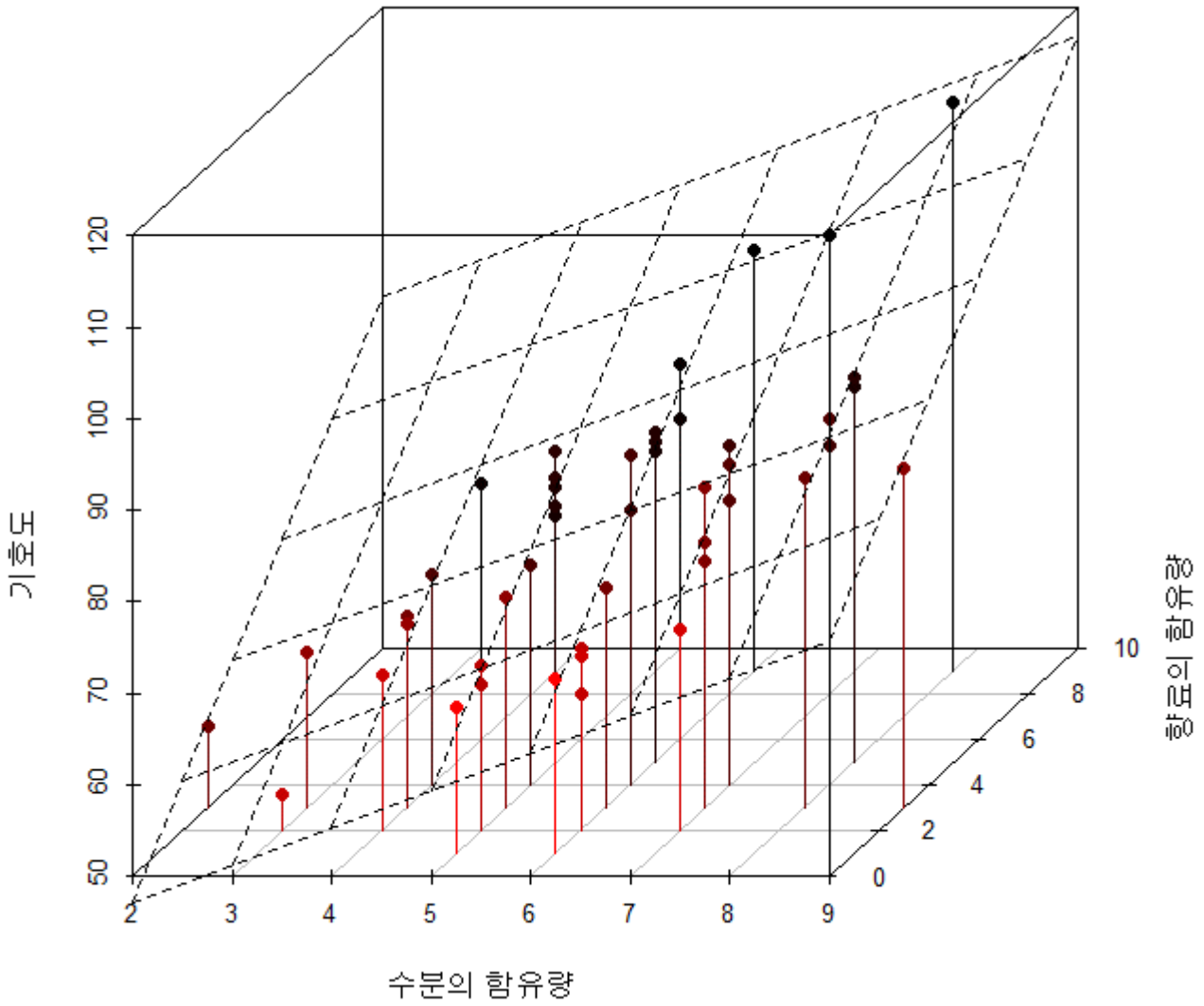
예제 2

- "cosmetic.txt " 는 어느 화장품의 기호도(Y)와 수분의 함유량(X1), 향료의 함유량(X2)의 관계를 분석하기 위하여 6명의 소비자를 랜덤으로 추출하여 얻은 자료이다.
- 그래프와 회귀분석을 통하여 변수간의 관련성을 파악하시오.

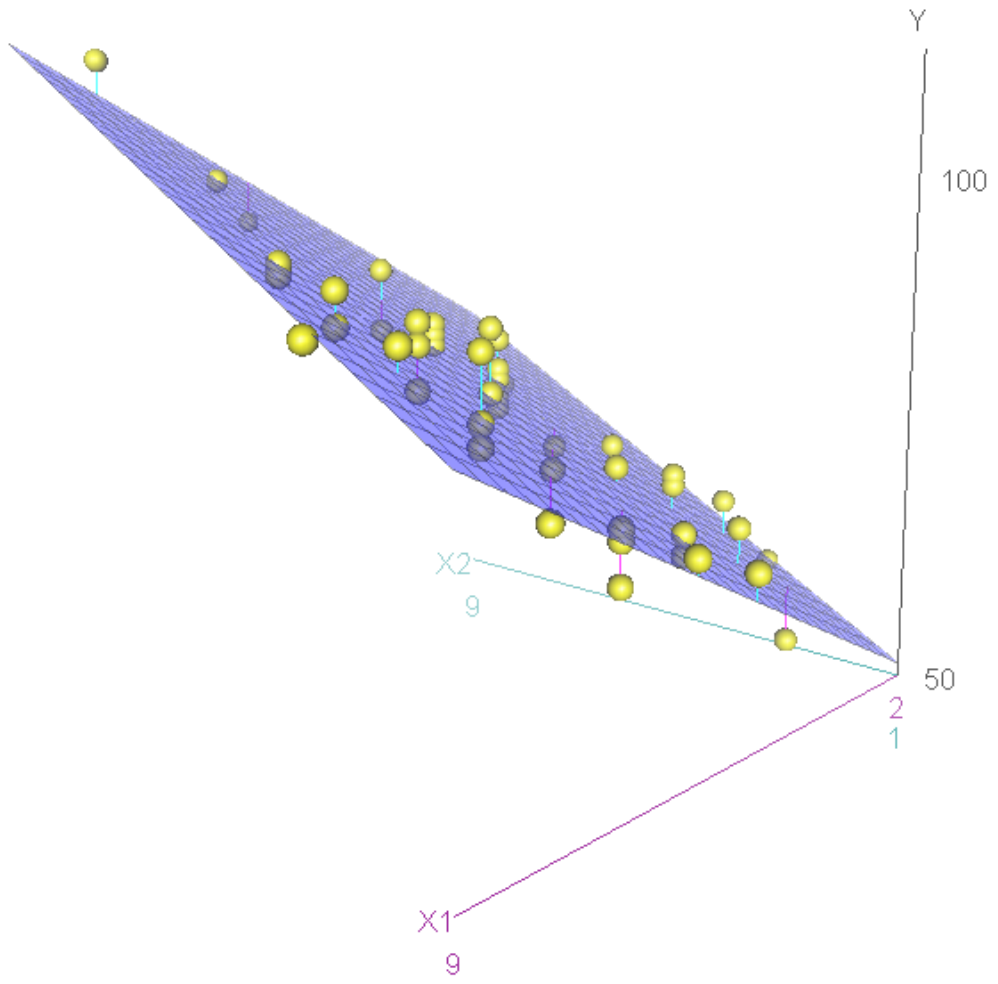
– 변수간의 관련성을 살펴보기 위한 상관도를 작성하시오.



– 회귀면을 포함한 3차원 차트를 작성하시오.

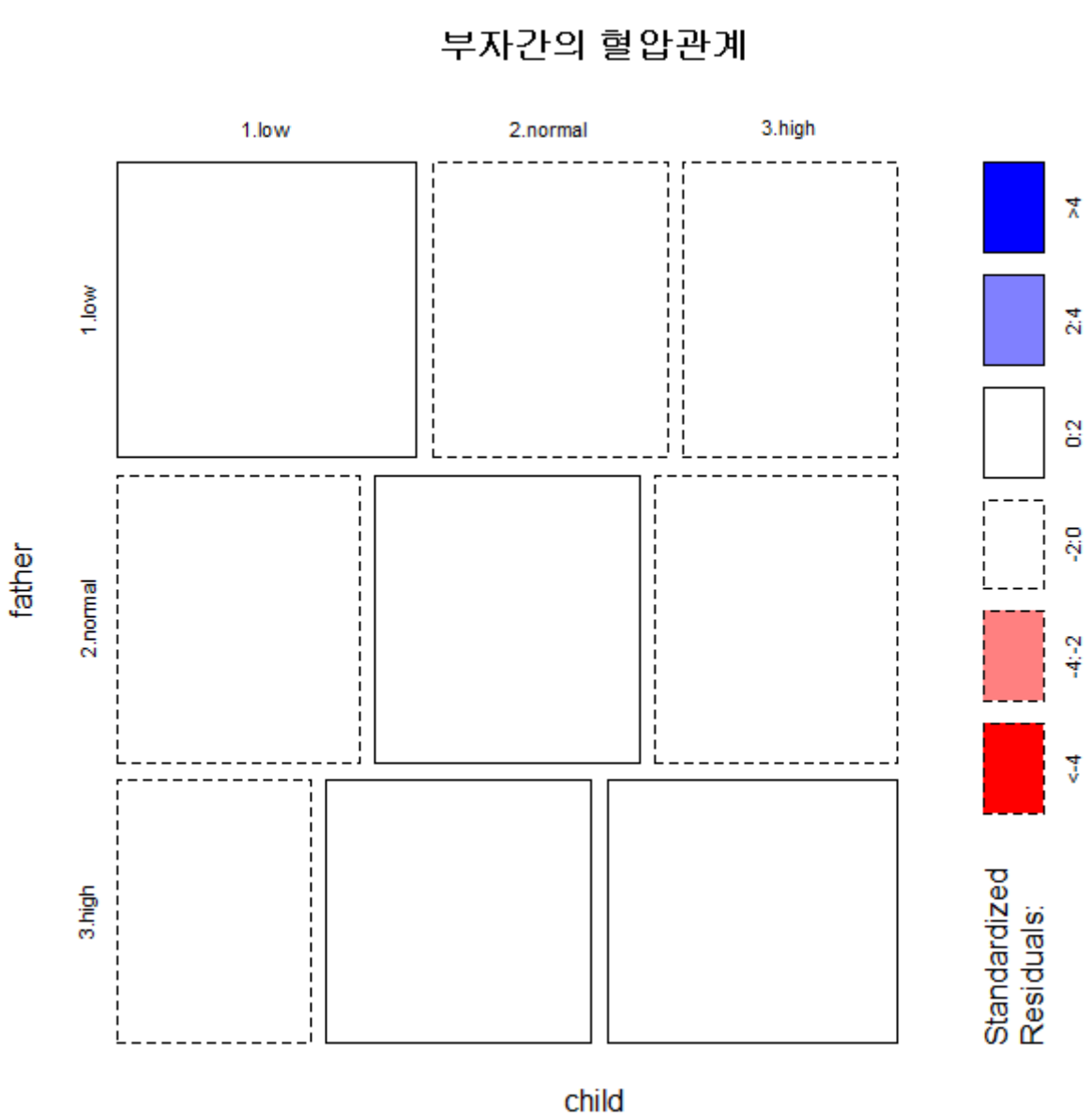


– 회귀면을 포함한 움직이는 3차원 차트를 작성하시오.



예제 4

- 혈압은 심장질환의 원인이 되는 것으로 알려져 있다. 이에 S의대의 K교수는 아버지의 혈압과 자녀의 혈압간에 유전적인 관련성이 있는지를 조사하여 심장질환 예방과 관련된 방안을 연구하고자 한다. 이를 위해 K교수는 100명의 초등학생을 랜덤으로 추출하여 이들의 아버지의 혈압과 함께 측정하여 아래와 같은 분할표를 얻었다. 아버지 혈압과 자녀의 혈압간에 관련성이 있는지를 유의수준 α=0.05에서 검정을 수행하고 모자이크 그림을 작성하여 독립성이 만족하는지 점검하시오.


부자간의 혈압관계
1.low
2.normal
3.high
father
1.low
2.normal
3.high
child
Standardized Residuals:
>4
2:4
0:2
-2:0
-4:-2
<-4

예제 4

- 학습목표
  - 서울도시철도공사에서 제공받은 2010~2013년 지하철 역별 승하차 정보 데이터를 바탕으로 탑승객 수를 역, 노선, 연도, 월별로 자료를 정리하는 기법을 습득하고, 탑승객 기준 상위 10개 역을 추출하여 이를 시각화 하는 방법을 학습한다.
  - 구글지도를 활용하여 추출된 자료를 시각화하는 기법을 학습하며, 도출된 이상자료가 해당날짜에 어떤 이벤트에 기인하여 나타나는지 뉴스자료 검색을 통해 그 연관성을 파악하는 기법을 학습한다.

- 본 실습에서는 R을 활용하여, 각 변수의 수준별 자료의 합, 평균 등 요약값을 계산하는 기법을 익히고, 자료분석에 필요한 데이터프레임을 생성하는 방법을 학습하며, ggplot2 패키지를 이용하여 추출된 정량정보의 시각화를 구현하는 기법을 익힌다.
- 구글맵을 R에서 활용하는 기법을 익혀 위도 및 경도 좌표계를 이용하여 지하철역의 위치를 시각화하고, 정량정보를 지도에 맵핑하는 기법을 활용하는 기법을 소개한다.

- **활용 데이터 소개**
  - **subway.xlsx : 2010년 1월부터 2014년 7월까지 서울지하철역 및 시간대별 승하차 인원수 정보를 제공**

변수명	설명
station	역코드
stat_name	역명
income_date	일자
on_tot	당일 해당역의 총 탑승인원 수
on_xx	당일 해당역의 xx시간대의 탑승인원 수 (xx는 05부터 24까지)
off_tot	당일 해당역의 총 하차인원 수
off_xx	당일 해당역의 xx시간대의 하차인원 수 (xx는 05부터 24까지)

- **subway_latlong.xlsx : [서울 열린데이터 광](http://data.seoul.go.kr/openinf/sheetview.jsp?infId=OA-118)에서 제공하는 지하철 노선별 역이름 및 위치정보(위도, 경도) 자료 및 각 역의 노선명을 제공**

변수명	설명
STATION_CD	역코드
STATION_NM	역명
LINE_NUM	호선
FR_CODE	외부코드(외국인의 경우, 역명보다 역번호로 문의하는 경우가 많음)
CYBER_ST_CODE	사이버스테이션(환승역의 경우 마스터가 되는 노선의 전철역코드)
XPOINT	X좌표
YPOINT	Y좌표
XPOINT_WGS	X좌표(WGS)
YPOINT_WGS	Y좌표(WGS)

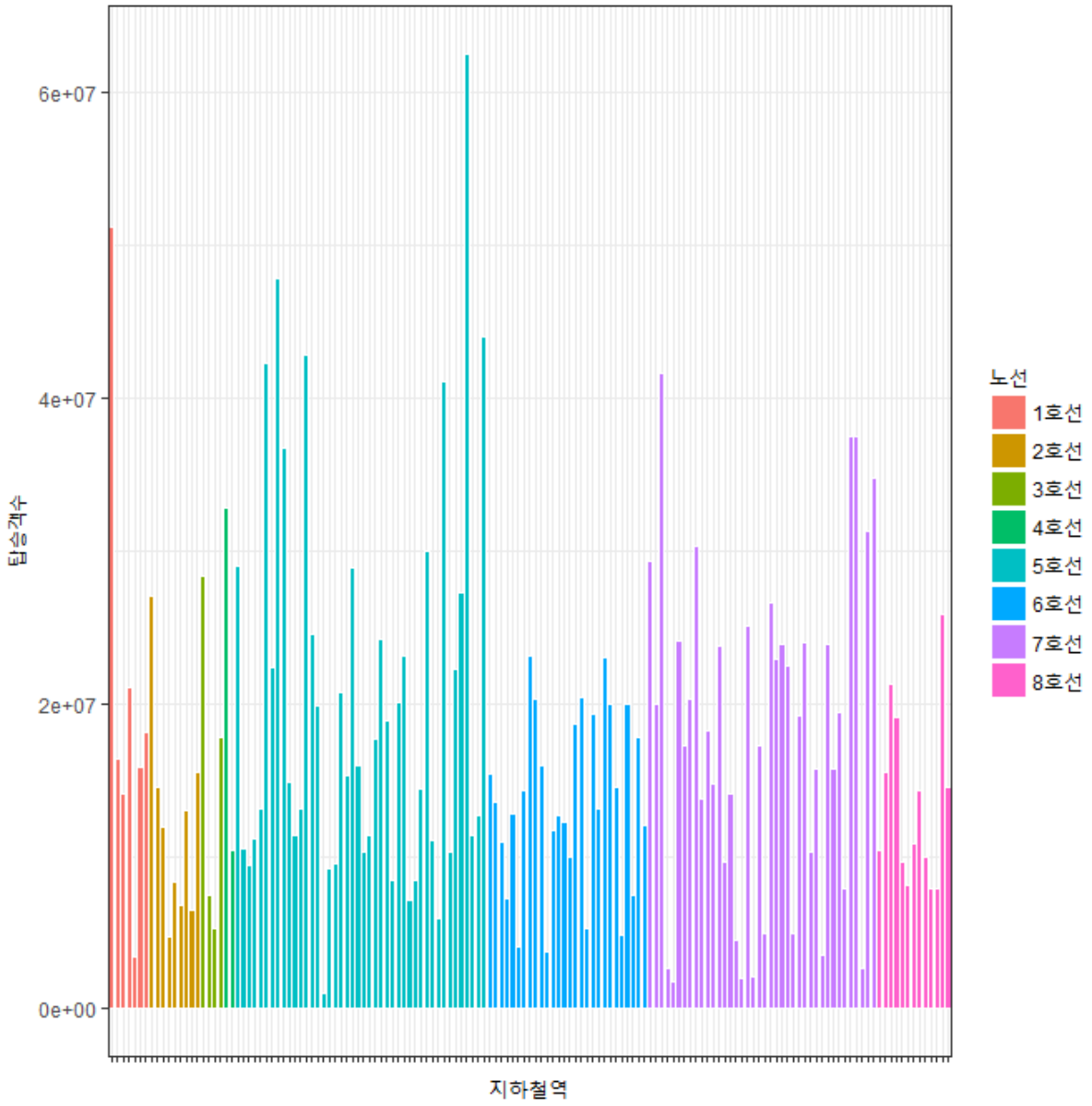
데이터 로딩 및 전처리 과정

- 본 장에서는 csv 파일 포맷으로 저장된 파일을 R에서 불러들여 데이터 객체를 생성하고, 분석에 활용 가능하도록 자료를 정리하는 과정을 학습

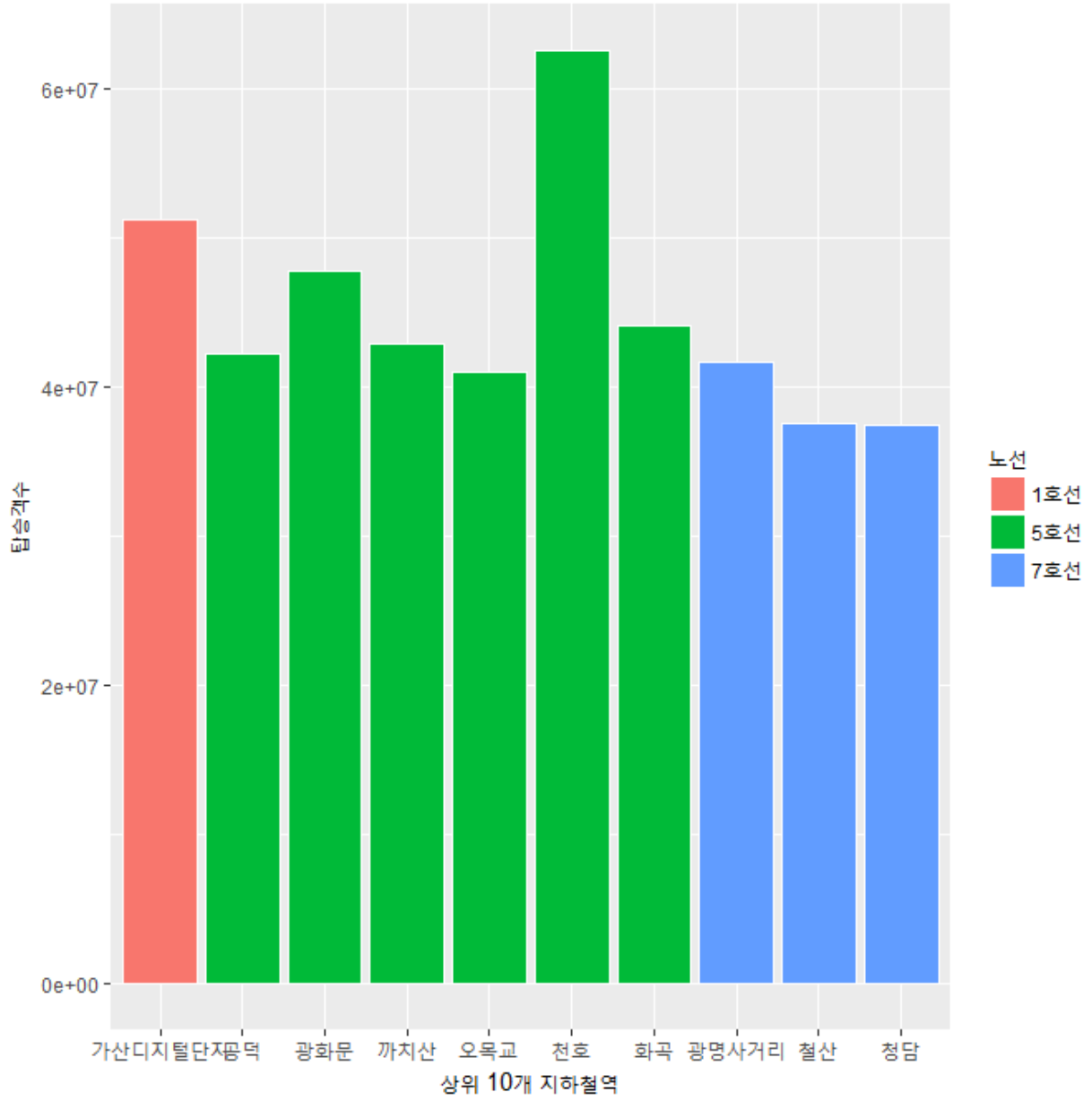
지하철역별 수치요약 및 시각화

- 본 장에서는 역별, 시간별 탑승객 수의 합산을 수행하여, 탑승객수 기준 상위 혹은 하위 역들을 파악하고, 시간별 탑승객수의 추이를 살펴보는 기법을 시각화 과정을 통해 살펴보는 방법을 학습한다.

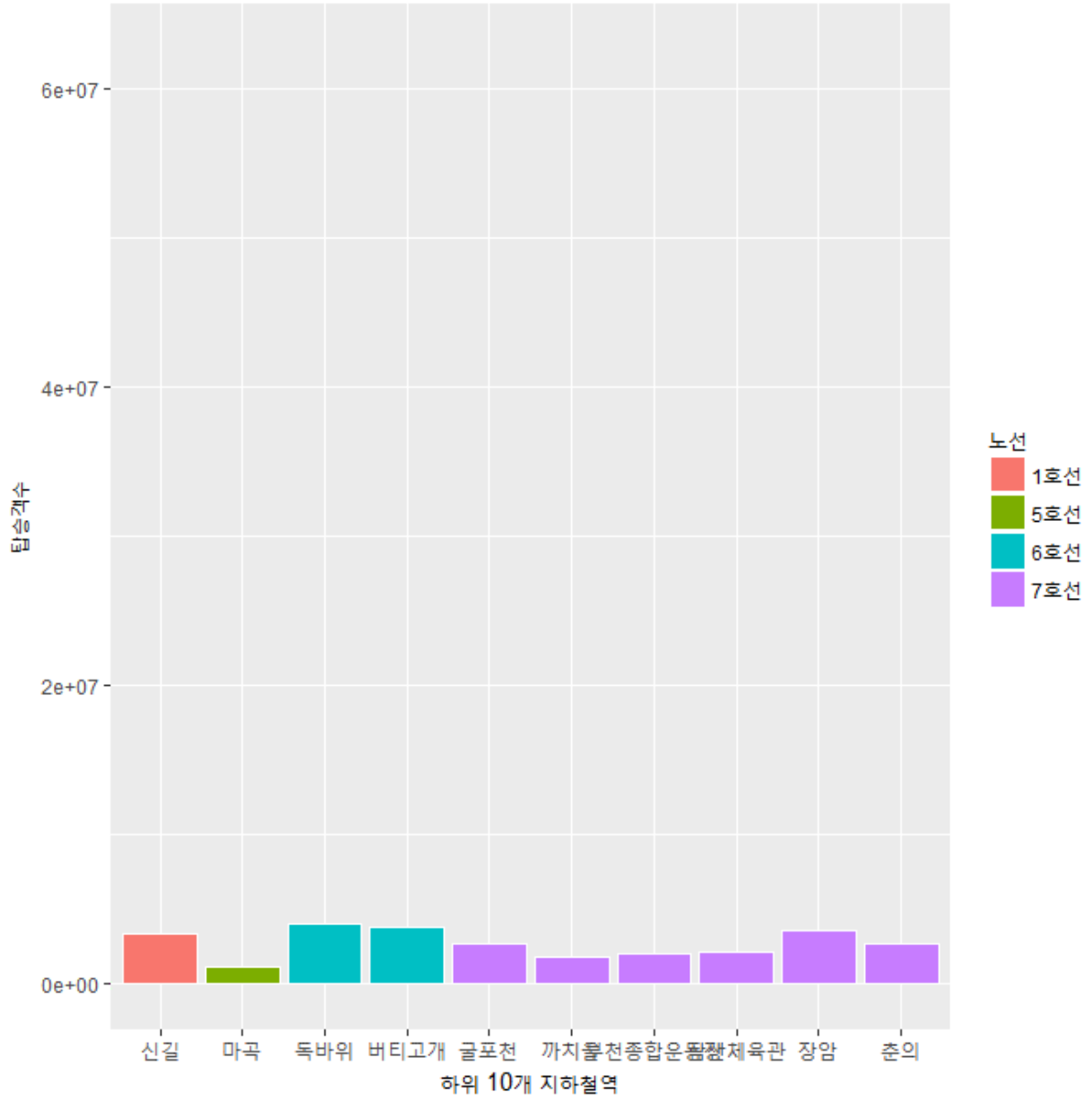
– 노선 및 역별 누적탑승객 시각화



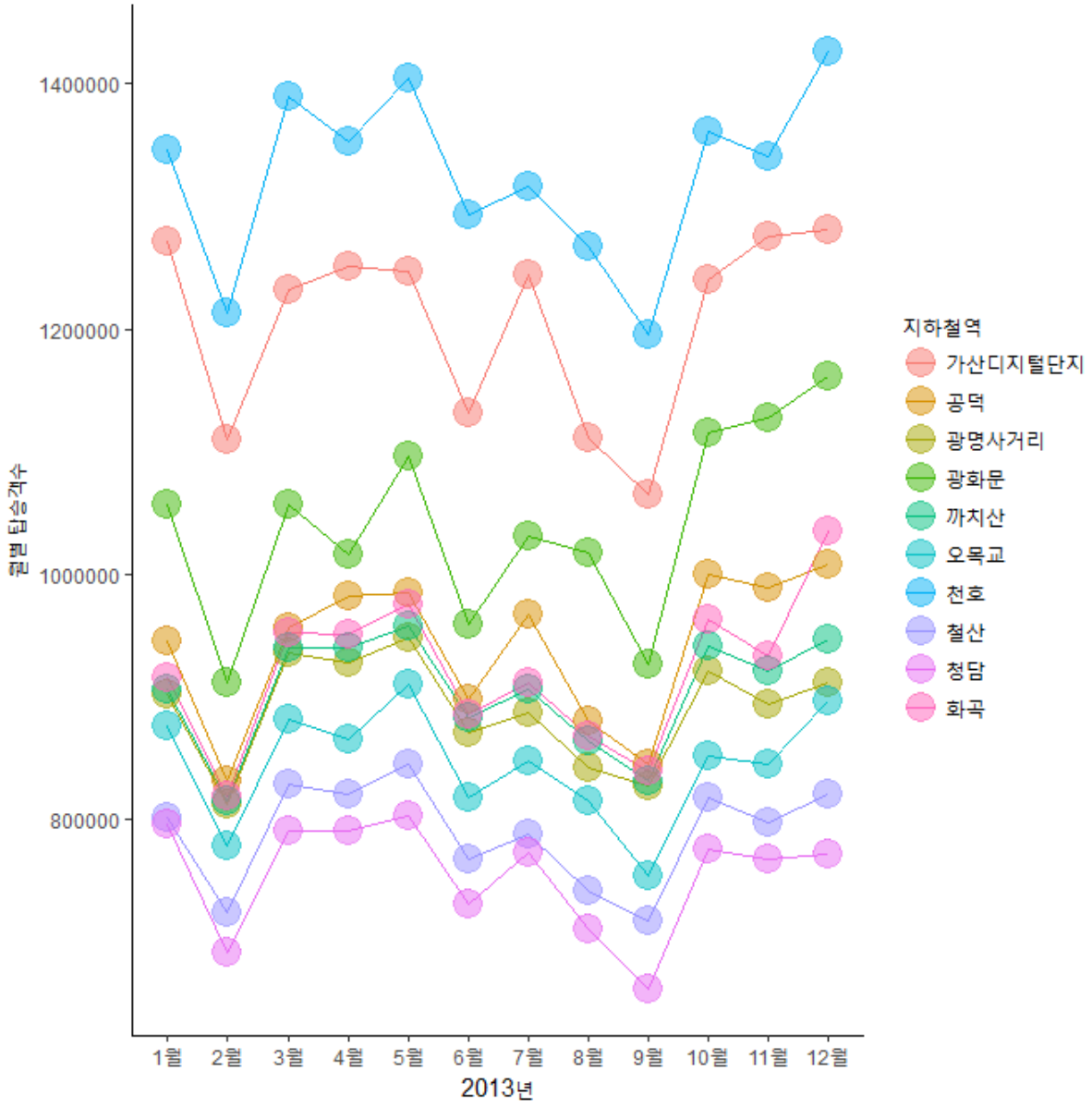
– 총누적탑승객 상위 10개역 탑승객수



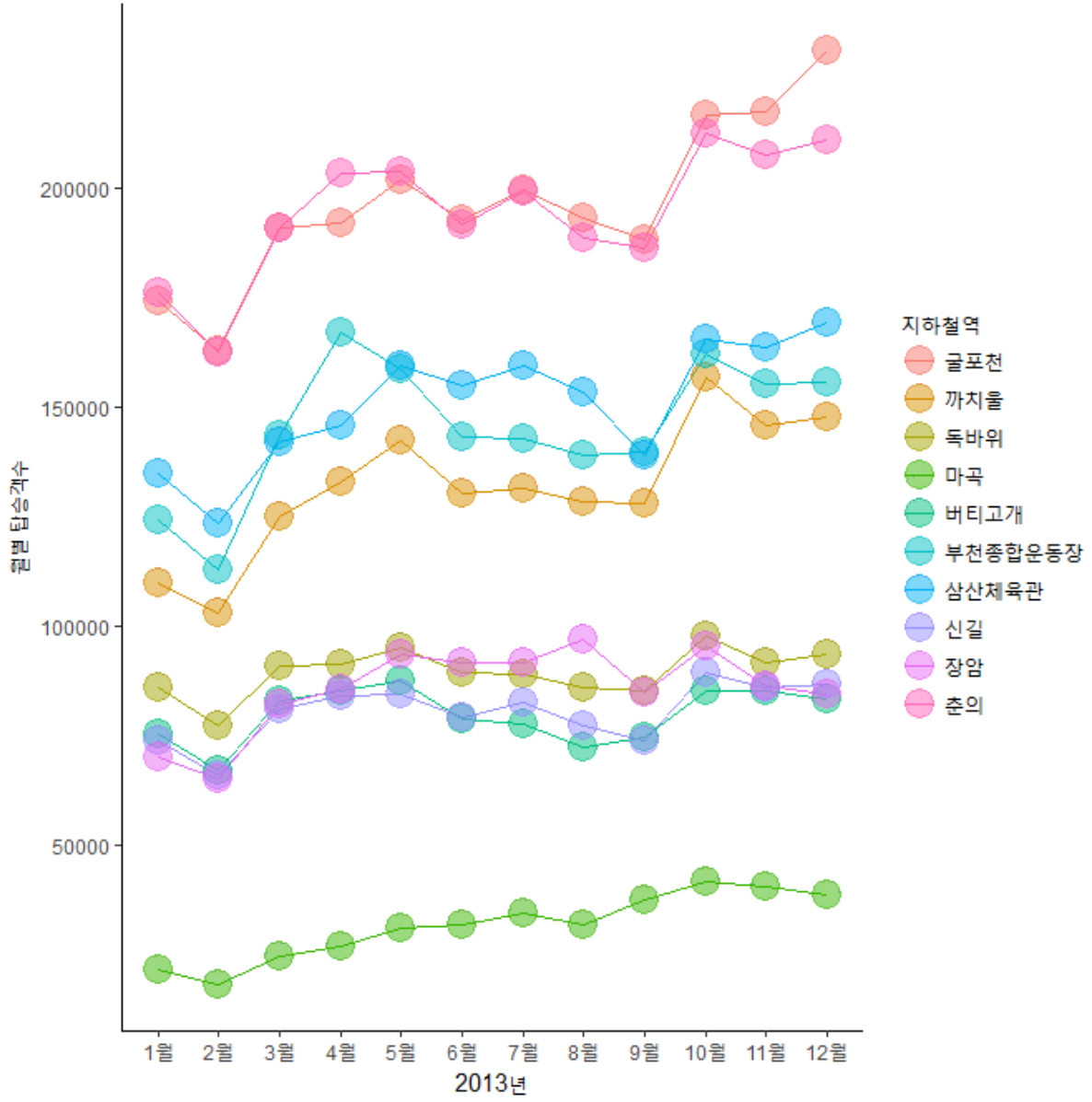
– 총누적탑승객 하위 10개역 탑승객수



– 2013년 상위 10개역 월별 탑승객



– 2013년 하위 10개역 월별 탑승객

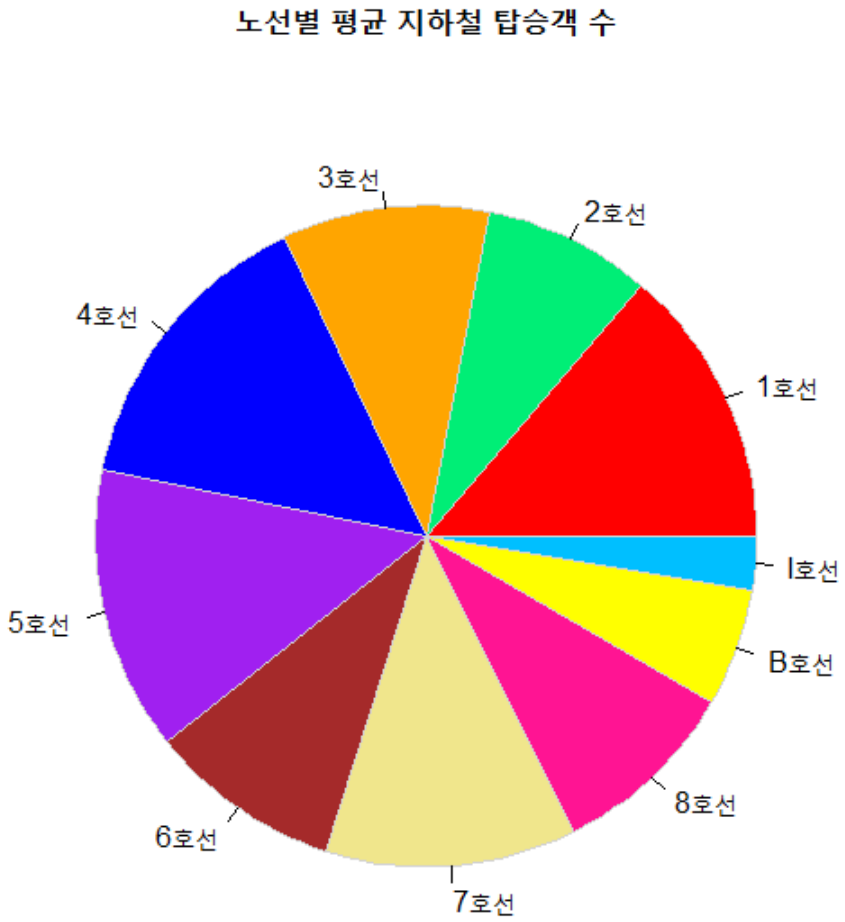


– 2013년 하위 10개역 월별 탑승객

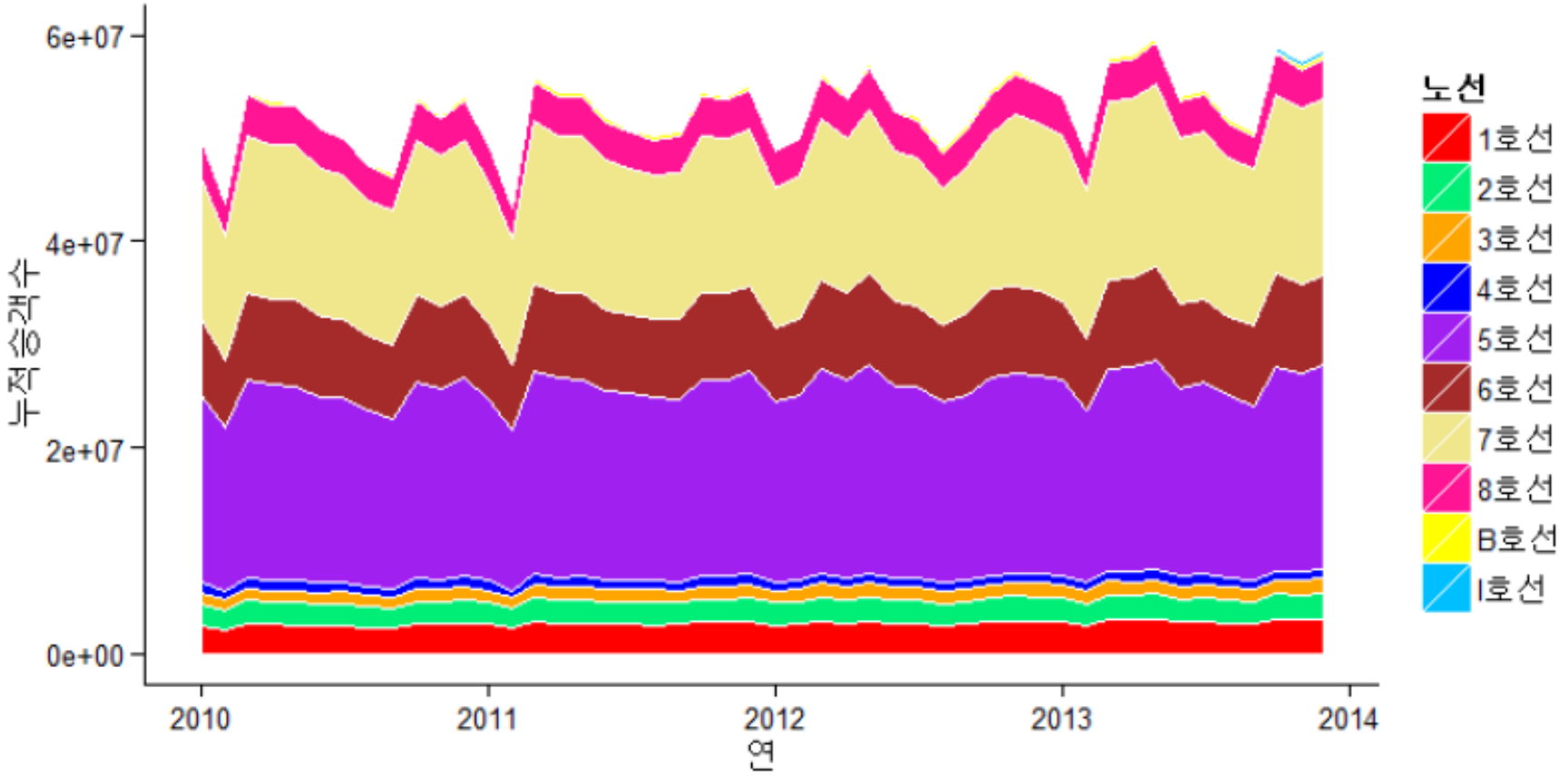
지하철노선별 수치요약 및 시각화

- 각 노선의 모든 역을 포함하고 있지 않으므로 누적 승객수를 계산하는 것은 전체 노선사용자를 나타내는데 적절치 않다.
- 따라서 노선에 포함되는 역들의 평균 탑승객수를 계산하여 이를 각 노선의 탑승객수 비교에 이용한다.

– 노선별 평균 지하철 탑승객수



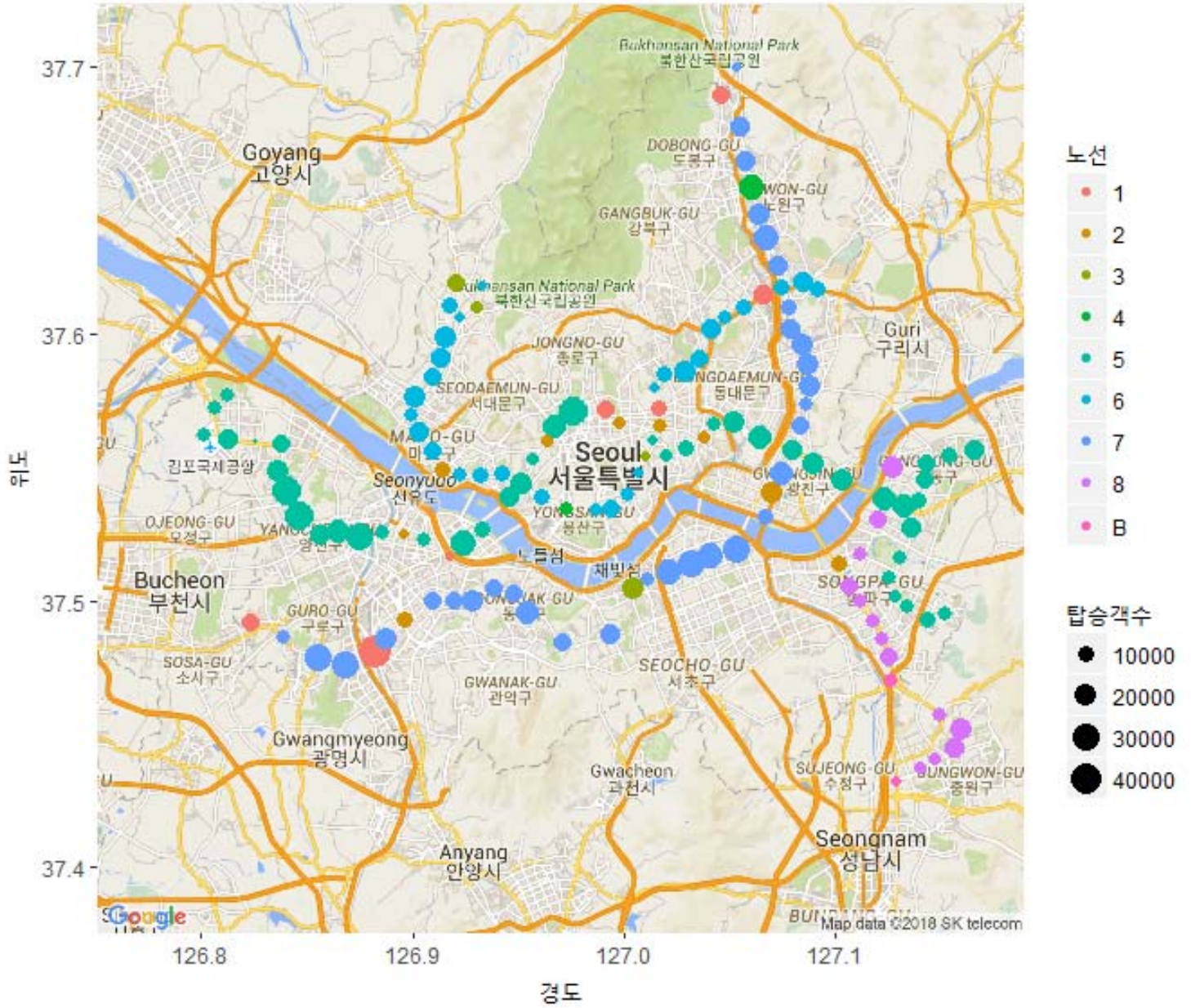
– 노선별 누적 탑승객수


6e+07
4e+07
2e+07
0e+00
누적 승객수
2010
2011
2012
2013
2014
연
노선
1호선
2호선
3호선
4호선
5호선
6호선
7호선
8호선
B호선
I호선

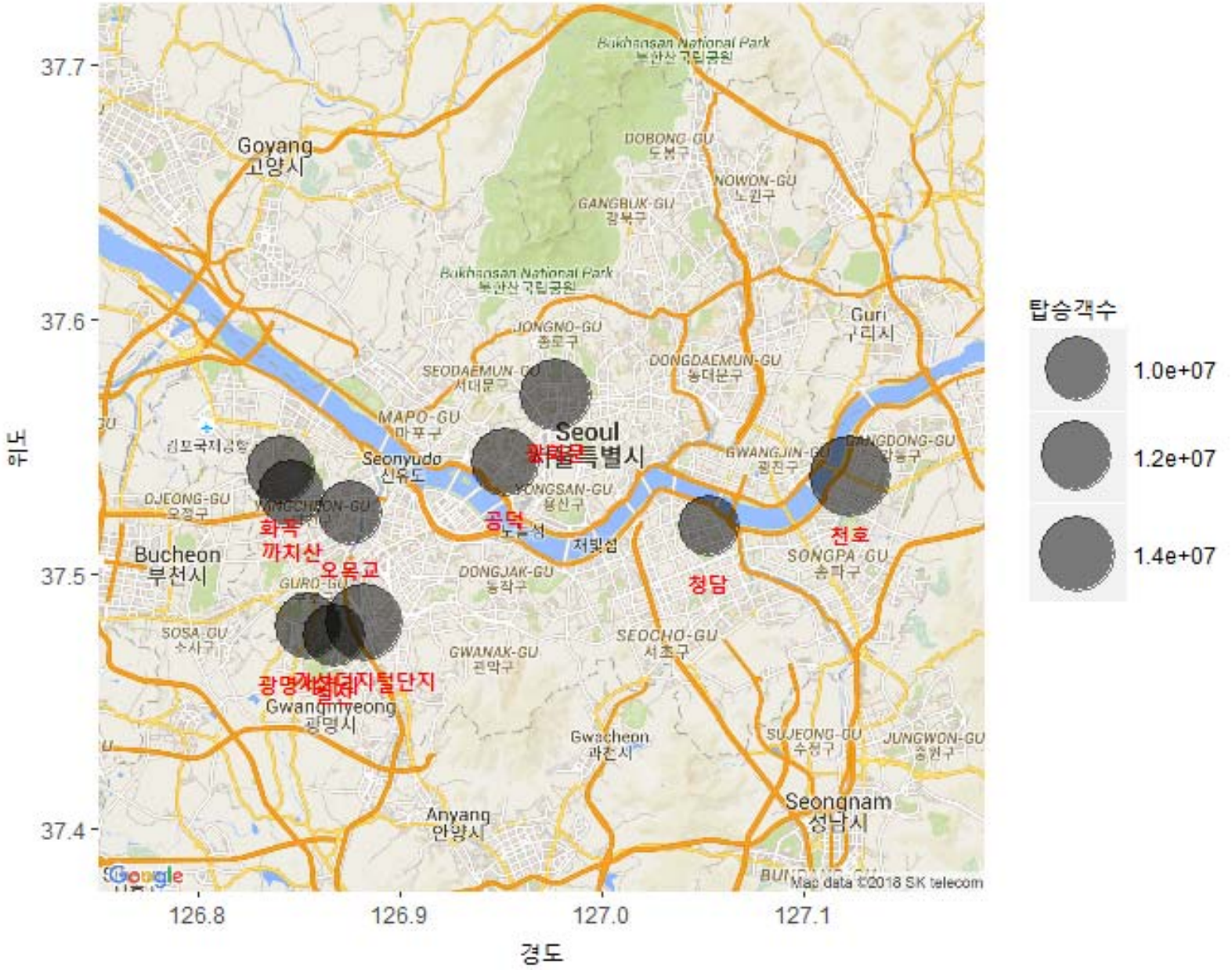
구글맵을 활용한 지도 맵핑

- 수도권지하철정보를 시각화하기 위해 서울역(위도 37.55, 경도 126.97)을 중심한 인근지역을 구글로부터 다운받아 이를 시각화한다.
- 아래는 2012년 5월 8일 하루동안 탑승한 인원을 각 역별로 수치화하고 이를 지하철역 위치에 크기에 비례하는 원으로 표현하고 있다.

– 각 역별 2012년 5월 8일 하루 탑승한 인원



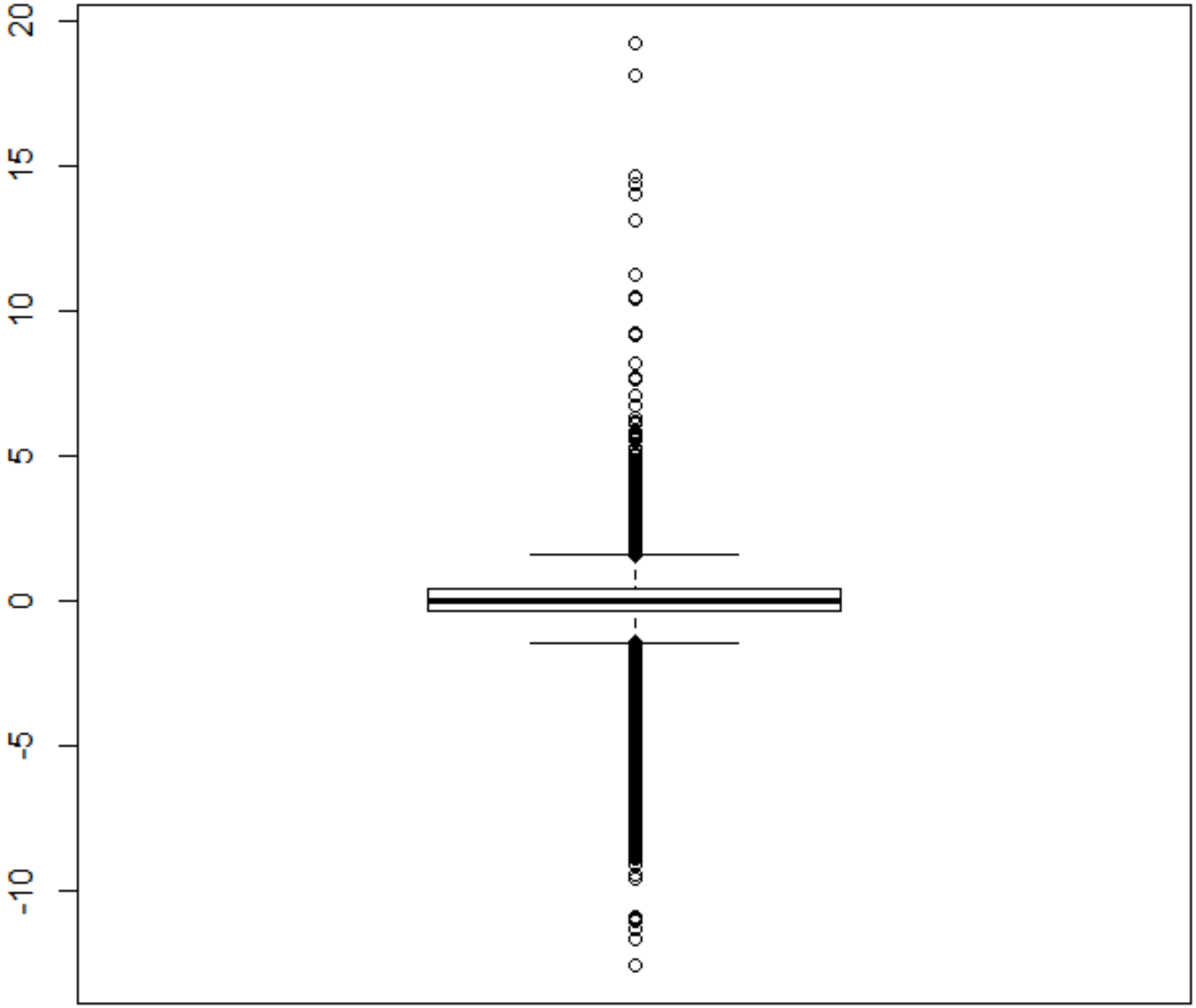
– 상위 10개역의 2013년도의 탑승객 수



회귀모형

- 회귀분석은 반응변수를 다수의 설명변수의 값에 의해 예측하는 모형을 제공한다.
- 본 절에서는 탑승객수를 지하철역, 지하철노선, 달, 요일의 요인에 의해 예측하는 모형을 적합한다. 탑승객수를 예측하기 위한 모형으로 회귀분석은 그리 정교하다고 하긴 어렵다.
- 하지만, 전반적인 패턴은 단순한 회귀분석을 통해 도출 가능하며 예측모형으로 널리 쓰인다.

– 잔차분석



– 2013년도 "상암월드컵"역의 탑승객 시간순 자료 시각화