데이터과학

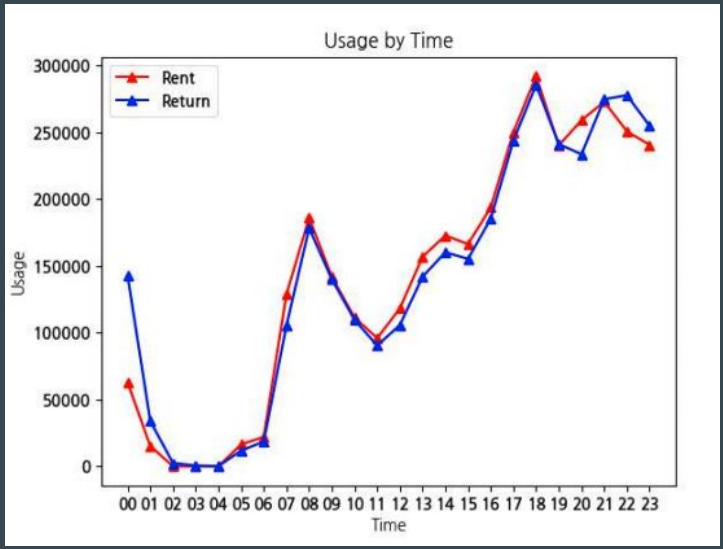
...

타슈데이터 - R

지난주...

- 제출률 : 32/48
  - 66.7%
- 질문
  - Classroom : 3개
  - E-mail : 10개
- 주로 실수하는 과정
  - test코드 수정
    - column명
    - 대문자를 소문자로..(RENT_STATION -> rent_station)
    - 임의의 column명으로 (NO, NAME..)
- 마감전 덜완성하고 마감 이후 제출한 과제가 존재한다면 뒤에 제출한 과제로 채점진행

GOOD-1



분석

1시간 간격의 사용량을 시각화한 결과, 7시 ~ 9시와 17시~19시가 사용률이 두드러지고 오전, 오후 시간보다 저녁 이후의 사용량이 월등히 많다.

앞서 짚은 두 시간대는 출/퇴근 및 등/하교 시간대로 장기간 이용이 아닌 한시간 내에 목적지까지 가기위한 인원으로 보인다.

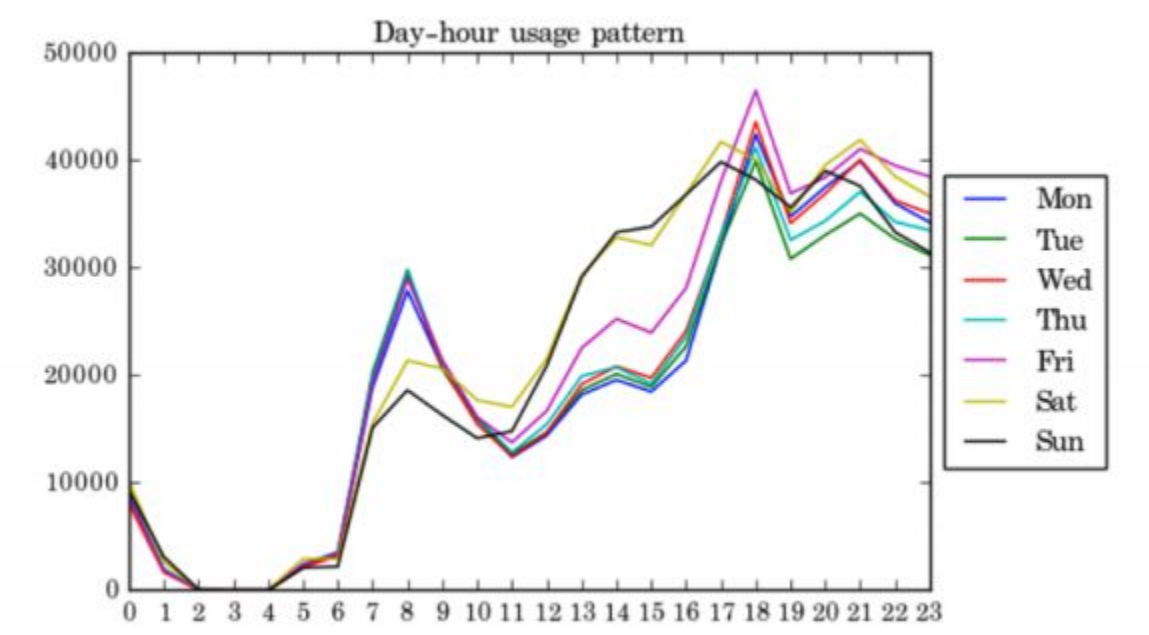
일과 후에 여가생활을 위한 사용으로 높은 사용률을 보이며, 00시~01시의 반납 사용률이 대여 사용률보다 높다.
이는 저녁시간대에 실컷 사용하고 자정 즈음에 반납하는 인원이 많다고 보여진다.

새벽 시간에는 대다수의 인원이 취침중이므로 사용률이 바닥을 보인다.

시간별 + 요일별과 지역(구)별 사용률을 통해 스테이션 및 자전거의 수를 효율적으로 조절할 수 있을것으로 보인다.

GOOD-2

- 주어진 문제 이외에 추가적인 Feature를 고려



기타

- 과제 채점 점수
- https://docs.google.com/a/cs-cnu.org/spreadsheets/d/1qR_Alrma9eIqFlvpNW8zMRgoWB7RzoLXtIoRP32ELbI/edit?usp=sharing

목표

- Data Analysis Process 따라하기 2
- 타슈데이터 재사용
  - 공공데이터 포털
- R 을 사용한 타슈데이터 분석
  - 인기 정류장 TOP 10 with **Google Map & ggplot**
  - 인기 경로 TOP 10 with ggplot

Data(재사용)

- station.csv

	번호	키오스크 번호	구별	명칭	위치	주소	거치대	좌표
1	1	1	유성구	무역전시관입구(택시승강장 앞)	엑스포다리 맞은편	유성구 도룡동 3-8	14	36.374325,
2	2	2	유성구	대전컨벤션 센터 앞	둔산대교 맞은편	유성구 도룡동 4-19	20	36.374472,
3	3	3	서구	한밭수목원(정문입구)	한밭수목원 내	서구 만년동 396	19	36.369855,
4	4	4	서구	초원아파트104동부근(버스정류장)	초원아파트 104동앞 쪽문 육교 버스정류장 앞	서구 만년동 401	12	36.368192,
5	5	5	서구	둔산대공원 입구(버스정류장)	한밭수목원에서 평송수련원 가는길 버스정류장 앞	서구 둔산동 1521-10	13	36.365034,
6	6	6	서구	백합4가 앞(농협앞)	백합아파트 상가 농협 버스정류장 앞	서구 월평2동 266	12	36.362304,
7	7	7	서구	정부청사 입구(대덕대로)	둔산 시외버스터미널 버스정류장 앞	서구 둔산동 920-2	13	36.361665,
8	8	8	서구	정부청사 입구(샘머리)	둔산 고속버스터미널 버스정류장 앞	서구 둔산동 1518	12	36.361794,
9	9	9	서구	황실아파트앞(성룡초교 앞)	성룡초교 정문 버스정류장앞	서구 월평동 304	12	36.361392,
10	10	10	서구	만년동 KBS 부근(기업은행 앞)	초원아파트 102동 육교 건너편	서구 만년동 300	12	36.369207,
11	11	11	서구	누리아파트앞(후문버스정류장)	누리아파트 후문과 무지개아파트 사이 버스정류장 앞	서구 월평3동 301	12	36.358995,
12	12	12	서구	정부청사역 앞(4번 출구)	삼성생명 앞	서구 둔산2동 949-1	13	36.357945,
13	13	13	서구	삼천중학교 앞	수정타운 아파트 1동 버스정류장 앞	서구 둔산2동 911	12	36.358597,
14	14	14	서구	둔산 하이마트 앞	둔산 이마트 맞은편 하이마트 버스정류장 앞	서구 둔산2동 962	20	36.355558,
15	15	15	서구	둔산 홈플러스 앞	법원 버스정류장 앞	서구 둔산동 1380-5	13	36.355591,

Data(재사용)

- tashu.csv

	RENT_STATION	RENT_DATE	RETURN_STATION	RETURN_DATE
1	43	2.01301e+13	34	2.013010e+13
2	97	2.01301e+13	*NA*	2.013010e+13
3	2	2.01301e+13	10	2.013010e+13
4	106	2.01301e+13	105	2.013010e+13
5	4	2.01301e+13	4	2.013010e+13
6	21	2.01301e+13	105	2.013010e+13
7	90	2.01301e+13	91	2.013010e+13
8	13	2.01301e+13	30	2.013010e+13
9	1	2.01301e+13	1	2.013010e+13
10	1	2.01301e+13	2	2.013010e+13
11	1	2.01301e+13	2	2.013010e+13
12	9	2.01301e+13	23	2.013010e+13
13	27	2.01301e+13	27	2.013010e+13

R로 데이터 분석

- 대여 정류소 출력 실습
  - RENT_STATION 의 TOP10 빈도수 출력

	rent_station	Freq
3	3	174801
8	8	52471
1	1	49886
4	4	40404
10	10	38894
5	5	34530
7	7	30616
6	6	26017
2	2	25670
9	9	21205

R 데이터 분석

- 자료형태 파악 후 생각

```
tashu_csv <- read.csv("tashu.csv")
station_csv <- read.csv("station.csv")

str(tashu_csv)
str(station_csv)
```

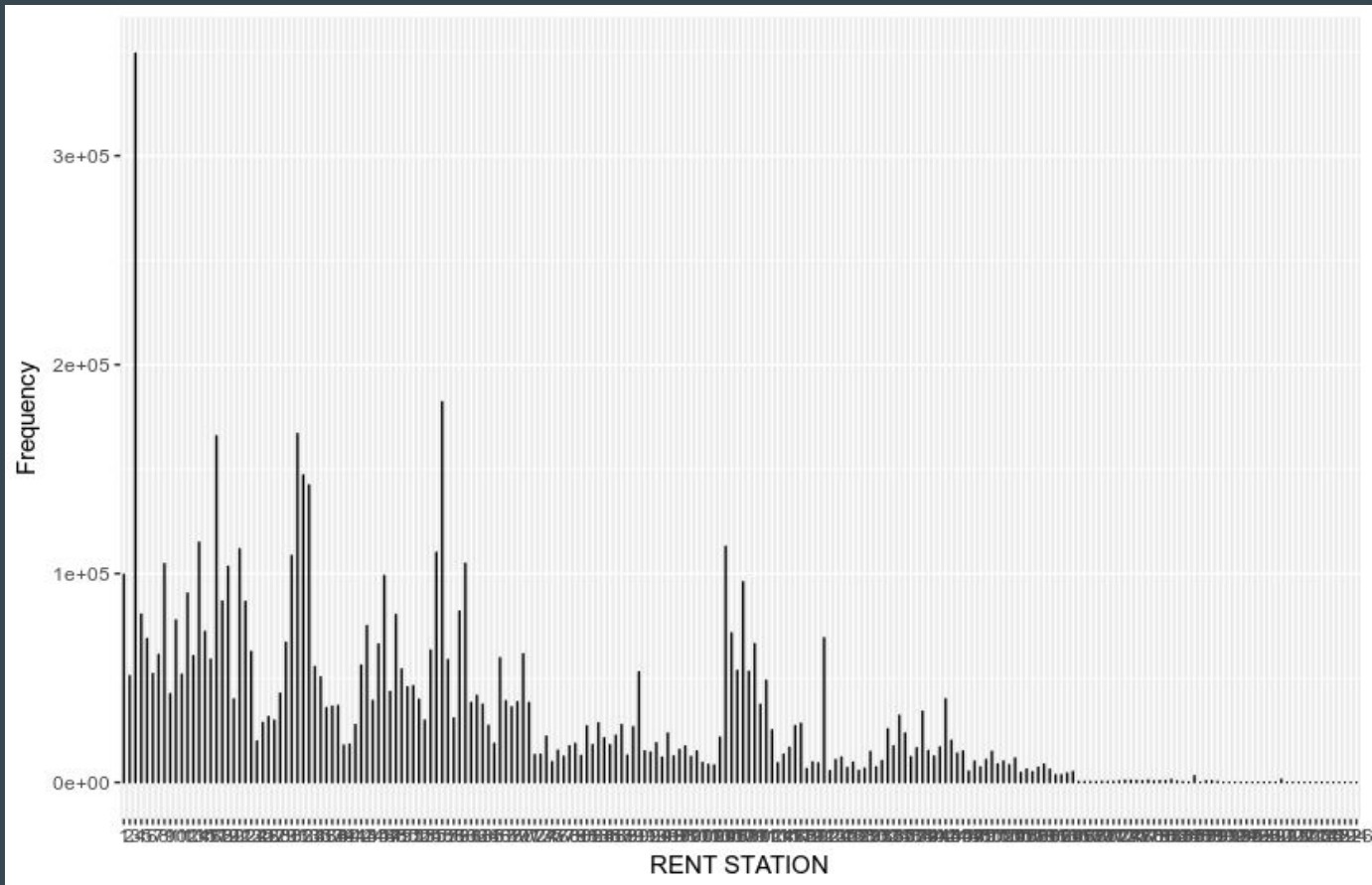
```
> str(tashu_csv)
'data.frame':   3404663 obs. of  4 variables:
 $ RENT_STATION  : int  43 97 2 106 4 21 90 13 1 1 ...
 $ RENT_DATE     : num  2.01e+13 2.01e+13 2.01e+13 2.01e+13 2.01e+13 ...
 $ RETURN_STATION: int  34 NA 10 105 4 105 91 30 1 2 ...
 $ RETURN_DATE   : num  2.01e+13 2.01e+13 2.01e+13 2.01e+13 2.01e+13 ...
> str(station_csv)
'data.frame':   144 obs. of  8 variables:
 $ 번호        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ 키오스크번호: int  1 2 3 4 5 6 7 8 9 10 ...
 $ 구별        : Factor w/ 5 levels "대덕구","동구",..: 4 4 3 3 3 3 3 3 3 3 ...
 $ 명칭        : Factor w/ 144 levels " 가람아파트앞",..: 95 87 72 128 23 96 121 55 144 91 ...
 $ 위치        : Factor w/ 114 levels "","CGV영화관 \n버스정류장 앞",..: 65 30 108 86 109 39 28 27 54 85 ...
 $ 주소        : Factor w/ 144 levels " 대덕구 법동 186",..: 83 84 49 50 40 55 46 39 60 48 ...
 $ 거치대      : int  14 20 19 12 13 12 13 12 12 12 ...
 $ 좌표        : Factor w/ 144 levels "36.303717, 127.457698",..: 126 127 120 117 108 101 95 98 94 119 ...
```

과제

1. 사용 정류장 TOP10 출력
    a. 막대 그래프
    b. Google Map
2. 사용패턴 TOP20 출력
    a. 막대 그래프

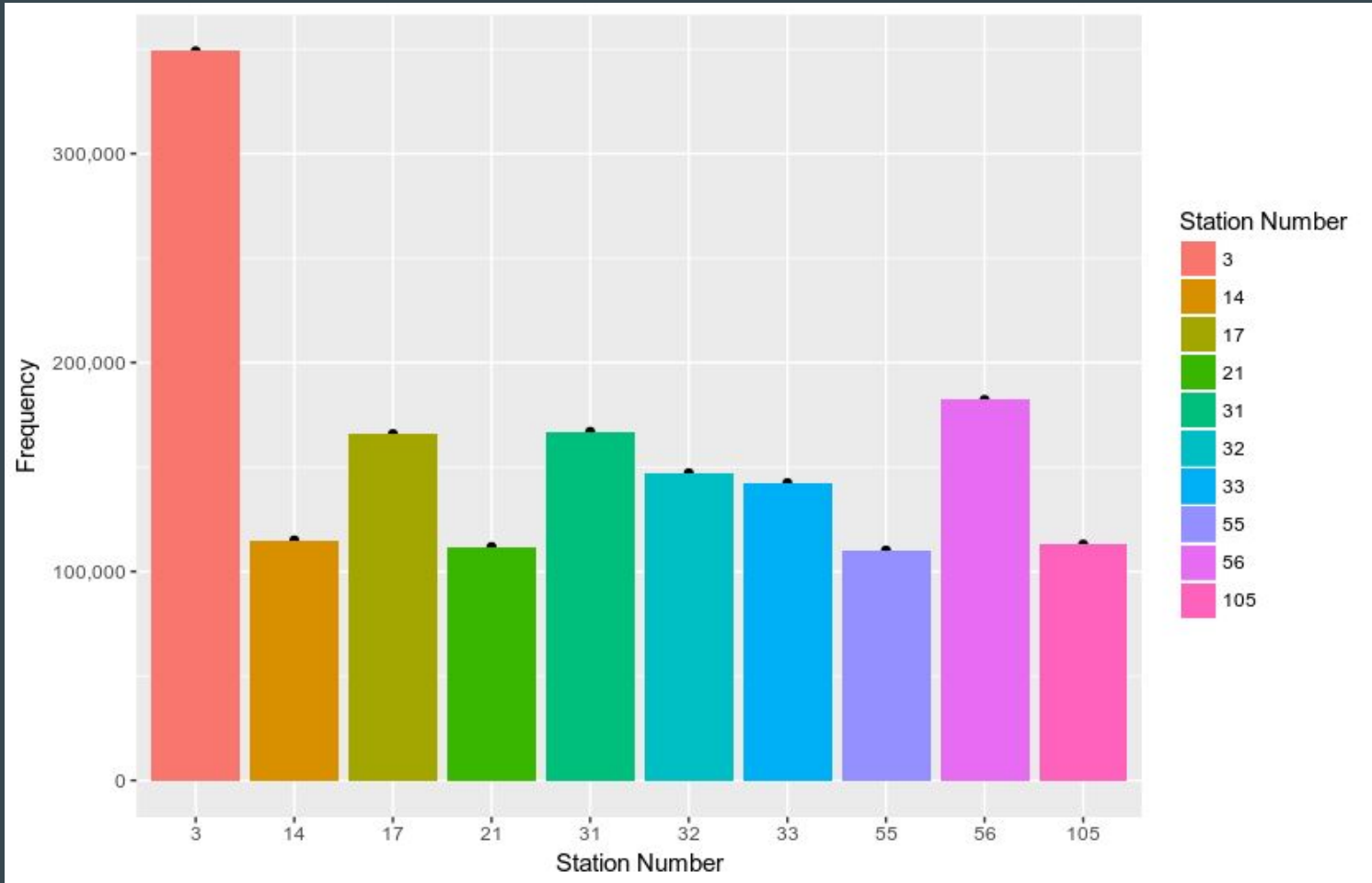
1. 정류장 사용 빈도

- RENT_STATION 과 RETURN_STATION을 합쳐서 빈도수 계산
- 전체 STATION 그래프 (복잡하니 TOP10으로 그려보자)(이건 과제 아님!)



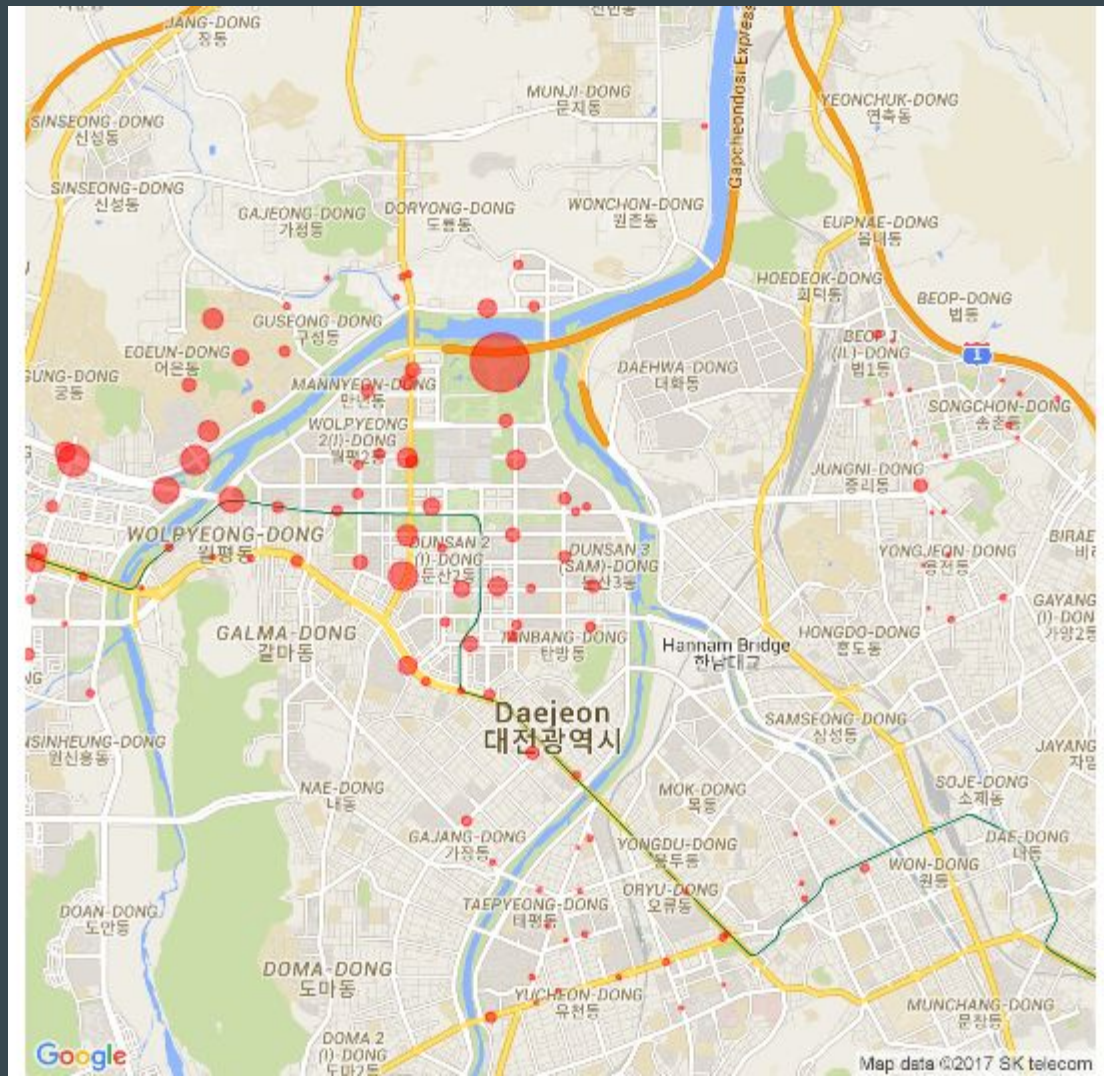
1. 정류장 사용 빈도

- ggplot2를 사용해 TOP10정류장 그리기



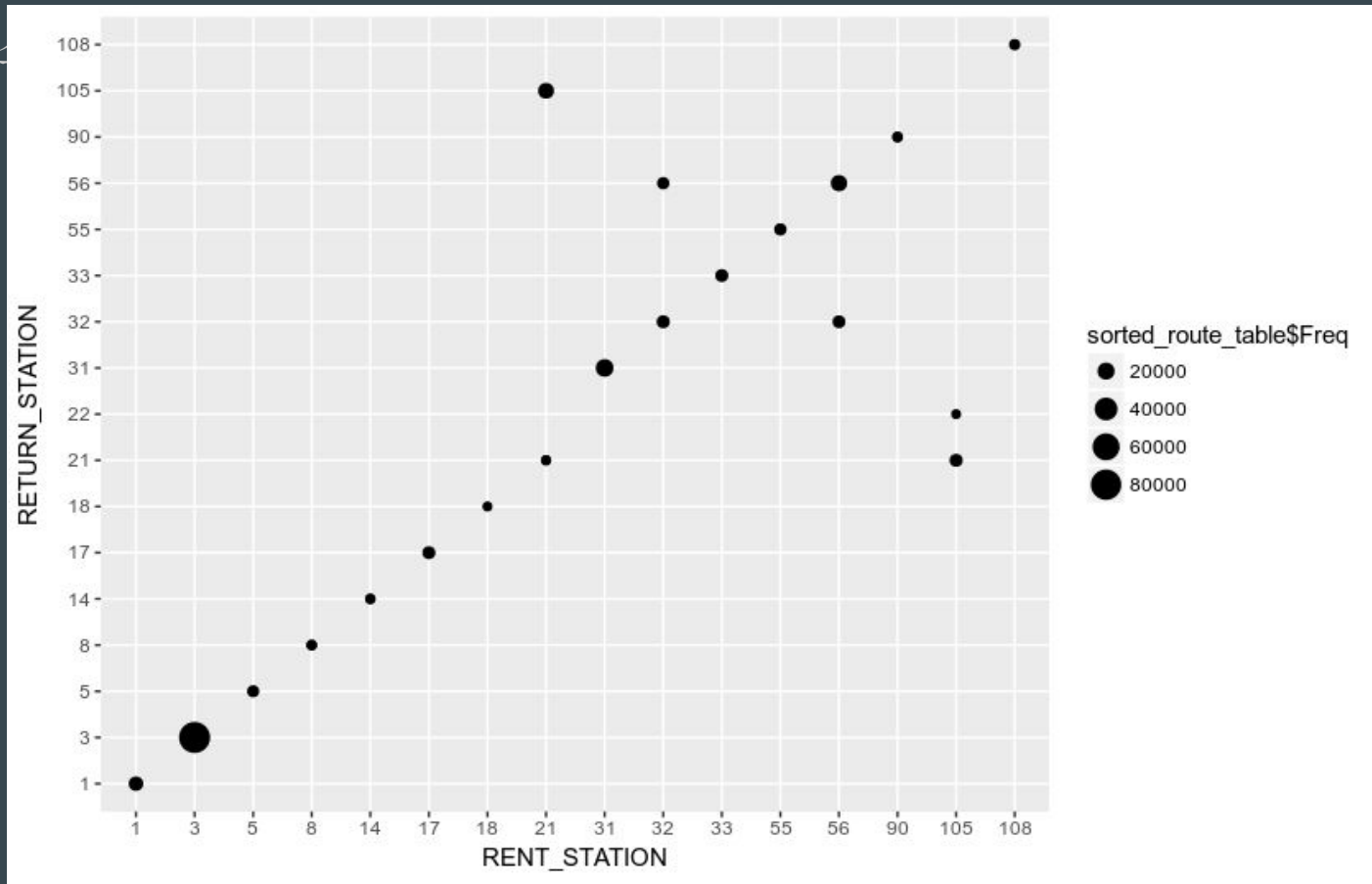
1. 정류장 사용 빈도

- ggmap 사용
- 각좌표로 Station 점찍기
- 사용빈도수가 점의 크기



2. 사용경로 빈도

- 빈도수가 높은 TOP 20 경로 그래프
- 가로 : 대여 정류소
- 세로 : 반납 정류소



과제 정리

- 총 3개의 그래프를 도출해내기
  - 인기 TOP10 정류장 막대그래프
    - RENT_STATION, RETURN_STATION 모두 고려
  - 사용빈도를 보기위한 지도
    - 빈도수가 곧 점의 크기
  - 인기 TOP20 경로 그래프

과제 제출 방법

- 과제 제출 기한 : 2017년 3월 22일 **오후 6시까지!**
- Google Classroom에 제출!
  - **24시간 경과시마다 20% 감점**
- 파일 제목 : **DS_학번_이름_주차.pdf, DS_학번_이름_주차.zip**
  - 보고서(PDF형태) : **HWP, DOC일경우 채점 안함**
    - 데이터 가공과정, 데이터 분석 과정, 코드 설명(스크린샷)
    - 결과(그래프)
  - 코드(파일 하나면 하나그대로, 여러개라면 ZIP형태로)
  - 코드파일과 보고서를 하나로 압축하지 말 것!
  - 파일 제목 및 형태 틀리면 -1점

질문 사항

- 방문
  - 606호 (데이터네트워크 연구실)
- 메일
  - dbgustlr92@cs-cnu.org
- Google ClassRoom
  - Good!