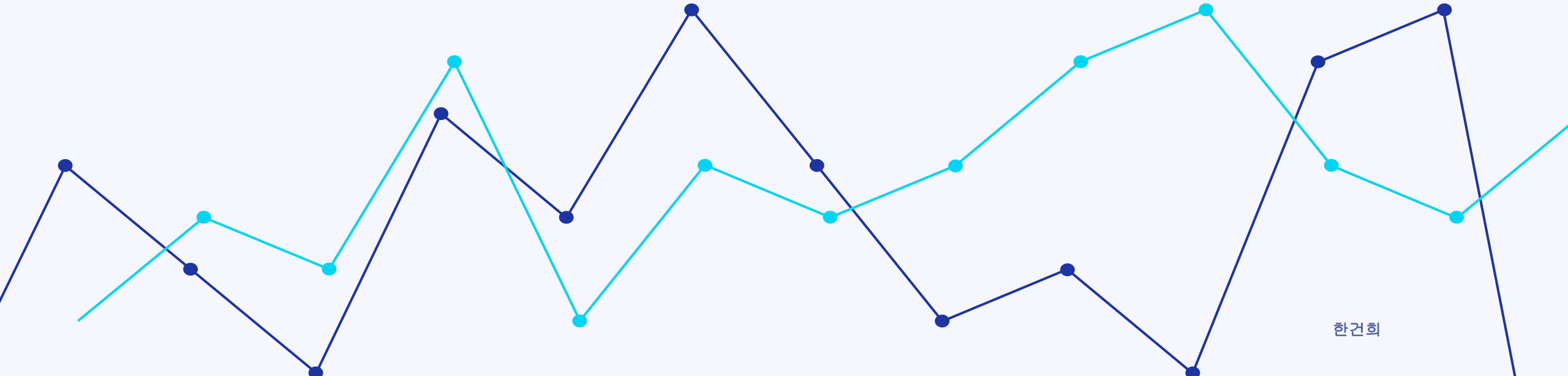
버스 지하철
데이터 분석 시각화

Analysis of Public Transportation Usage Status



한건희

Table of contents

Introduction

- 1949년 처음 버스 운행을 시작한 이후로 1970년대에 도심지역 교통난을 해결하기 위해 지하철(地下鐵) 등장
- 현재는 이동수단 이상의 가치
- '역세권'은 상권형성, 도시 개발, 부동산 가격에 영향

  -역세권이란 도시철도법에 따라 철도역과 인근 주변지역을 지칭

- 이용량을 통해 서울시 인구밀집도 유추 가능

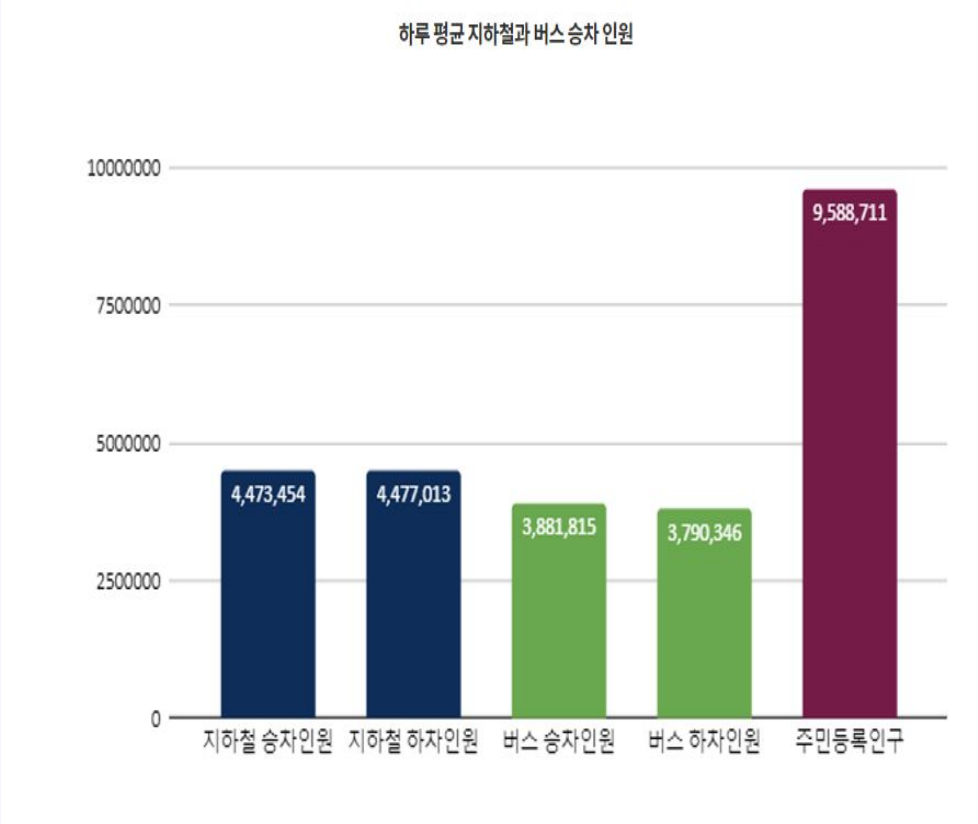
Purpose

- 서울시 주민등록인구 9백 5십만명인 점을 고려 매일 대중교통을 이용하는 사람들 규모
- 시간대/ 지역/ 기간에 따른 특징 분석

  ex) 퇴근 후 가장 많이 하차한 지역의 특징, 월별 승차,하차 인원의 특징(방학,여름휴가 등등)

- 코로나 이전과 이후 비교

> **인구통계학적 특징 분석**



출처: 과학기술정보 통신부

Data set

- 국토교통부-도시철도여객수송

(2015~2022년 서울,경기지역 버스/지하철 Data)

YYYYMM	S_ID	B_S	04-05in	04-05out	05-06in	05-06out	06-07in	06-07out	07-08in	07-08out	08-09in	08-09out	09-10in	09-10out	10-11in	10-11out	11-12in	11-12out
202208	1호선	동대문	561	16	9859	1842	8375	6305	13390	11046	17632	20315	16633	19979	16135	19311	17382	22249
202208	1호선	동묘앞	145	1	2799	1039	3456	4571	5920	8160	10055	17264	8698	16243	10088	19117	15536	24563
202208	1호선	서울역	573	19	8638	8274	12332	45706	39560	102779	63523	200999	51463	135243	53148	73952	66606	63535
202208	1호선	시청	39	0	2005	4665	3404	23606	6430	65621	8401	181920	10174	80351	12315	37637	18961	36074
202208	1호선	신설동	309	22	8586	2260	8758	9028	18458	22614	26047	54554	17926	30292	15728	20197	18093	17900
202208	1호선	제기동	357	4	5001	2038	8276	8838	21335	19703	31333	40232	22857	34466	26008	38818	33280	42490
202208	1호선	종각	54	4	3356	4382	3765	22971	5801	98968	9571	243599	11798	131015	16651	58516	25736	53733
202208	1호선	종로3가	118	10	3367	3149	3409	13161	4642	25201	8037	69020	12995	66338	19816	57735	30813	60906
202208	1호선	종로5가	38	2	1632	3635	2766	15329	5251	40866	8560	93100	12562	56834	20776	50437	30729	52138
202208	1호선	청량리(서	915	17	10286	4451	15174	21761	34968	17224	44626	34255	29439	30996	29565	36968	34584	40049
202208	2호선	강남	122	5	9527	11078	18061	51718	37772	148470	61996	302013	49820	305392	50089	145030	65762	103435
202208	2호선	강변(동서	18	0	8660	2269	26630	22755	76037	27501	110297	49117	80495	42037	63889	38757	56051	36715
202208	2호선	건대입구	355	13	15746	1658	21120	17389	50866	25152	89083	57608	59748	49836	35262	38915	34128	43500
202208	2호선	교대(법원.	37	1	2671	6919	12237	25299	24831	57742	35014	151298	28992	132481	27320	66399	32389	54223
202208	2호선	구로디지털	338	20	36294	5278	51355	20059	131269	85498	176670	244498	106045	141505	63060	56035	55410	41238

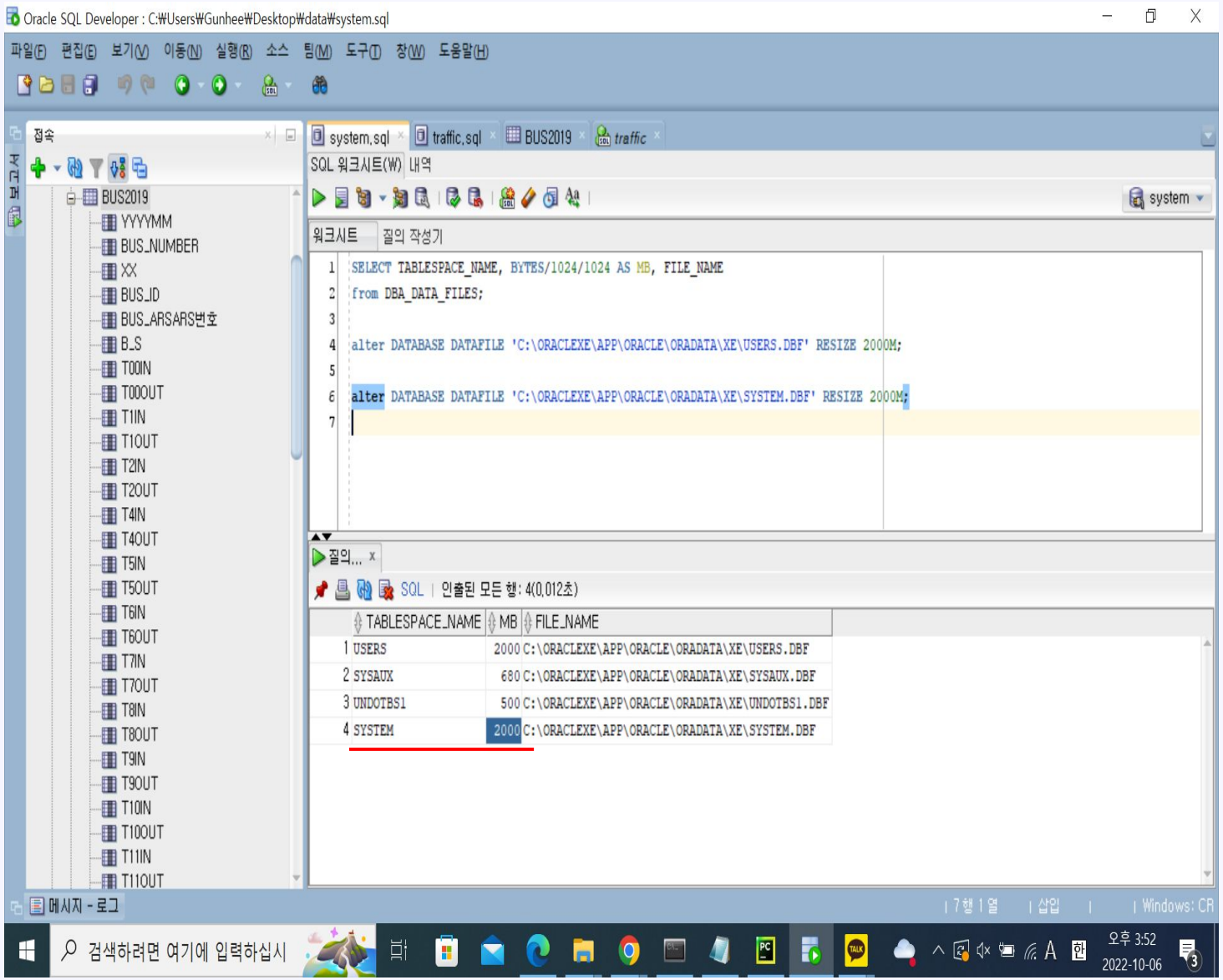
2015.01~2022.08 서울경기 시간대별 지하철 Data(53,884)

YYYYMM	bus_numb	xx	bus_ID	bus_ARS	B_S	t00in	t00out	t1in	t1out	t2in	t2out	t3in	t3out	t4in	t4out	t5in	t5out	t6in
202001	100	100번(하계	1E+08	1002	창경궁.서울	1	0	0	0	0	0	0	0	77	127	160	25	76
202001	100	100번(하계	1E+08	1003	명륜3가.성	0	0	0	0	0	0	0	0	113	180	163	300	60
202001	100	100번(하계	1E+08	1005	혜화동로터	0	2	0	0	0	0	0	0	106	89	100	310	8
202001	100	100번(하계	1E+08	1198	원남동	0	0	0	0	0	0	0	0	41	97	66	285	
202001	100	100번(하계	1E+08	1204	종로5가.효	19	34	1	0	1	0	0	0	0	0	0	0	
202001	100	100번(하계	1E+08	1205	종로5가.효	157	21	2	0	1	0	0	0	0	0	0	0	10
202001	100	100번(하계	1E+08	1212	종로5가	129	19	0	0	0	1	0	0	0	0	0	0	25
202001	100	100번(하계	1E+08	1219	통신대(이	73	52	0	1	0	1	0	0	0	0	0	0	12
202001	100	100번(하계	1E+08	1220	혜화역.마	214	80	14	0	2	1	0	0	0	0	0	0	18
202001	100	100번(하계	1E+08	1229	혜화역.동	408	117	15	4	6	0	0	0	0	0	0	0	83
202001	100	100번(하계	1E+08	1247	광장시장	1	0	0	0	0	0	0	0	14	85	31	153	3
202001	100	100번(하계	1.01E+08	2004	서울역버스	23	2	3	0	0	0	0	0	65	50	329	252	112
202001	100	100번(하계	1.01E+08	2007	서울역버스	152	54	3	2	1	0	0	0	0	0	9	39	39
202001	100	100번(하계	1.01E+08	2008	숭례문	0	0	0	0	0	0	0	0	4	79	54	219	3
202001	100	100번(하계	1.01E+08	2127	북창동.남대	0	1	0	0	0	0	0	0	18	64	121	128	

30배

2015.01~2022.08 서울경기 시간대별 버스 Data(1,722,051)

연구방법 - Oracle(11g_version)



한계 용량 존재(TableSpace)

> Data Size↑ → Data Import 실패

> Data Size(MB) 증가

연구방법 - Oracle(11g_version)

-- 버스 전체 데이터 확인 --

```
select count(*) from BUS2019;
select count(*) from BUS2020;
select count(*) from BUS2021;
select count(*) from BUS2022;
select count(*) from SUB;
```

-- 이상치 제거 --

```
delete from BUS2019 where  BUS_ARSARS번호 IN ('~');
delete from BUS2020 where  BUS_ARS IN ('~');
delete from BUS2021 where  BUS_ARS IN ('~');
delete from BUS2022 where  BUS_ARS IN ('~');
```

-- 서울시 이외 정류장 제거 --

```
delete
from BUS2019
where BUS_ARSARS번호 between 26000 and 99999;

delete
from BUS2020
where BUS_ARS between 26000 and 99999;

delete
from BUS2021
where BUS_ARS between 26000 and 99999;

delete
from BUS2022
where BUS_ARS between 26000 and 99999;
```

버스 ARS(버스 정류장 번호)에 일부 ”~” 有

→ 이상치 제거(10,067)

서울 이외의 정류장 제거

→ 총 1,722,051 → 1,581,113(제거수:140,938)

연구방법 - Oracle(11g_version)

```
-- 지하철 시간과 버스시간을 통일화 하기위해 버스시간컬럼 드랍.
ALTER TABLE BUS2019 DROP COLUMN T2IN;
ALTER TABLE BUS2019 DROP COLUMN T2OUT;
ALTER TABLE BUS2019 DROP COLUMN T3IN;
ALTER TABLE BUS2019 DROP COLUMN T3OUT;
ALTER TABLE BUS2019 DROP COLUMN 등록일자;
ALTER TABLE BUS2020 DROP COLUMN T2IN;
ALTER TABLE BUS2020 DROP COLUMN T2OUT;
ALTER TABLE BUS2020 DROP COLUMN T3IN;
ALTER TABLE BUS2020 DROP COLUMN T3OUT;
ALTER TABLE BUS2020 DROP COLUMN 등록일자;
ALTER TABLE BUS2021 DROP COLUMN T2IN;
ALTER TABLE BUS2021 DROP COLUMN T2OUT;
ALTER TABLE BUS2021 DROP COLUMN T3IN;
ALTER TABLE BUS2021 DROP COLUMN T3OUT;
ALTER TABLE BUS2021 DROP COLUMN 등록일자;
ALTER TABLE BUS2022 DROP COLUMN T2IN;
ALTER TABLE BUS2022 DROP COLUMN T2OUT;
ALTER TABLE BUS2022 DROP COLUMN T3IN;
ALTER TABLE BUS2022 DROP COLUMN T3OUT;
ALTER TABLE BUS2022 DROP COLUMN DATA_DAY;
ALTER TABLE SUB DROP COLUMN T02_03IN;
ALTER TABLE SUB DROP COLUMN T02_03OUT;
ALTER TABLE SUB DROP COLUMN T03_04IN;
ALTER TABLE SUB DROP COLUMN T03_04OUT;
ALTER TABLE SUB DROP COLUMN 작업일자;
```

버스 02~04시,작업일자 Columns Drop

연구방법 - Oracle(11g_version)

```
create TABLE SUB_TAMP AS
select YYYYMM, S_ID, B_S, T00_01IN,T00_01OUT,T01_02IN,T01_02OUT
,T04_05IN,T04_05OUT,T05_06IN,T05_06OUT,T06_07IN,T06_07OUT, T07_08IN,T07_08OUT,
T08_09IN,T08_09OUT,T09_10IN,T09_10OUT,T10_11IN,T10_11OUT,T11_12IN,T11_12OUT,T12_13IN,T12_13OUT,T13_14IN,
T13_14OUT,T14_15IN,T14_15OUT,T15_16IN,T15_16OUT,T16_17IN,T16_17OUT,T17_18IN,T17_18OUT,
T18_19IN,T18_19OUT,T19_20IN,T19_20OUT,T20_21IN,T20_21OUT,T21_22IN,T21_22OUT,T22_23IN,
T22_23OUT,T23_24IN,T23_24OUT,GU_ID
from SUB;
```

- Bus,Subway columns match
- Oracle 11g 이하 기존 Table 복사 후 새로운 Table 생성 → 기존 Table 삭제

연구방법 - Python

- 결측치,이상치 제거
  - ✔ 서울 이외의 지역 제거 종로구,중구,용산구,성동구,광진구,동대문구,중랑구,성북구,강북구,도봉구,노원구,은평구,서대문구, 마포구,양천구,강서구,구로구,금천구,영등포구,동작구,관악구,서초구,강남구,송파구,강동구(25개)만 추출
  - ✔ 사용년월, 대중교통 이상치 존재 → 제거
  - ✔ 당일 새벽 4시~다음날 새벽 1시까지 제외한 시간 제거(버스/지하철 비교)
    - → 사용년월,ARS,역명,시간대별(승차,하차)인원, 지역구,대중교통 총 49개 columns 추출

- 시간대별, 지역구, 연도, 월별, 대중교통(버스,지하철) Groupby 실행

- 승차, 하차인원의 문자 타입 변경(Float → Int)

연구방법 - Python

	Raw Data	결측치 제거	Modified Data	이상치 제거	Final Data
데이터 수	1,581,113	→	1,581,112	→	1,570,401
제거된 데이터 수		1개		10,711개	

```
raw_data_df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1581113 entries, 0 to 1581112
Data columns (total 49 columns):
 #   Column    Non-Null Count    Dtype
---  ------    --------------    -----
 0   사용년월      1581113 non-null  object
 1   ARS       1581112 non-null  object
 2   역명        1581112 non-null  object
 3   00시승차인원   1581112 non-null  float64
 4   00시하차인원   1581112 non-null  float64
 5   01시승차인원   1581112 non-null  float64
 6   01시하차인원   1581112 non-null  float64
 7   04시승차인원   1581112 non-null  float64
```

결측치 제거

```
modified_data_df1.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1581112 entries, 0 to 1581112
Data columns (total 49 columns):
 #   Column    Non-Null Count    Dtype
---  ------    --------------    -----
 0   사용년월      1581112 non-null  object
 1   ARS       1581112 non-null  object
 2   역명        1581112 non-null  object
 3   00시승차인원   1581112 non-null  float64
 4   00시하차인원   1581112 non-null  float64
 5   01시승차인원   1581112 non-null  float64
 6   01시하차인원   1581112 non-null  float64
 7   04시승차인원   1581112 non-null  float64
 8   04시하차인원   1581112 non-null  float64
```

이상치 제거
(서울 외의 지역 제거)

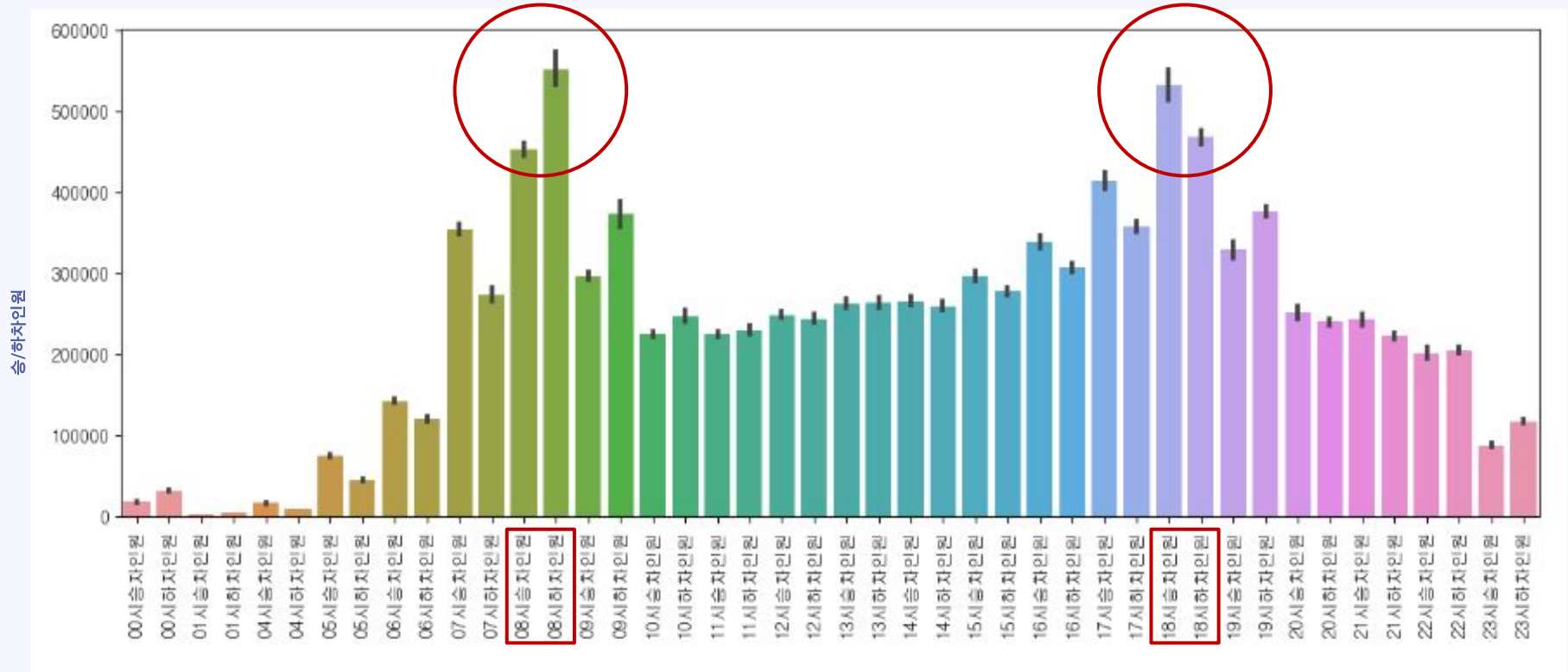
문자타입 변경
(실수 → 정수)

```
final_data_df.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1570401 entries, 0 to 1581112
Data columns (total 49 columns):
 #   Column    Non-Null Count    Dtype
---  ------    --------------    -----
 0   사용년월      1570401 non-null  int32
 1   ARS       1570401 non-null  object
 2   역명        1570401 non-null  object
 3   00시승차인원   1570401 non-null  int32
 4   00시하차인원   1570401 non-null  int32
 5   01시승차인원   1570401 non-null  int32
 6   01시하차인원   1570401 non-null  int32
 7   04시승차인원   1570401 non-null  int32
 8   04시하차인원   1570401 non-null  int32
```

Data Charts Infographics (Time)

시간대별 승/하차 총 인원

2019.01 ~ 2022.08



<서울 기준>
08시(출근)에 하차, 18시(퇴근) 승차 多

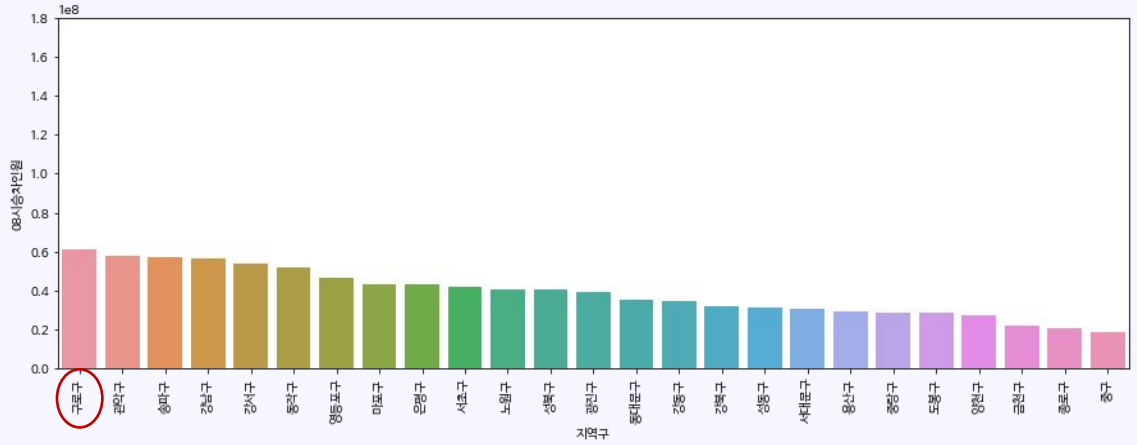
Data Charts Infographics (Region)

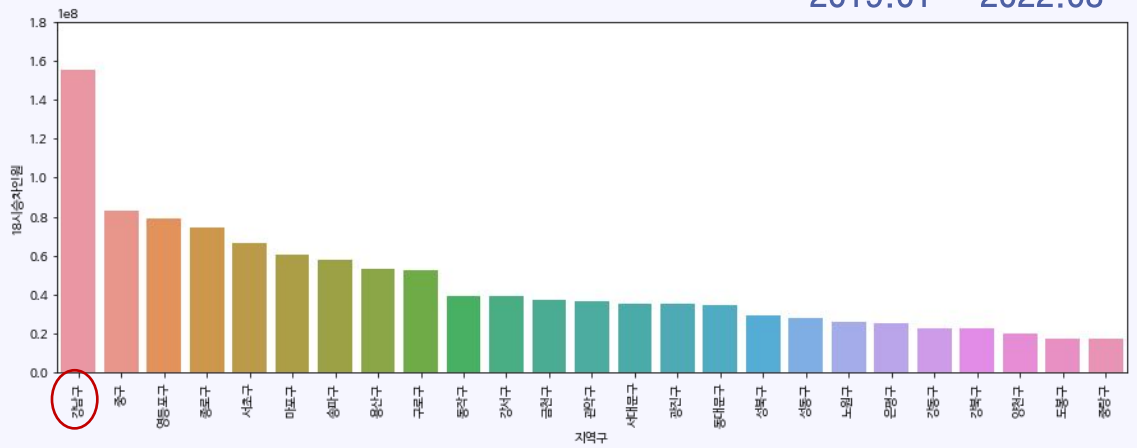
08시

18시

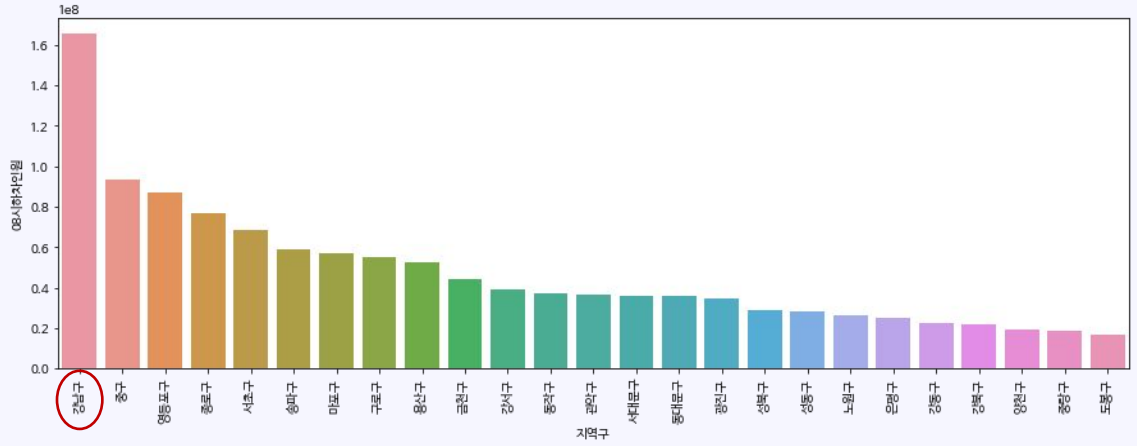
2019.01 ~ 2022.08

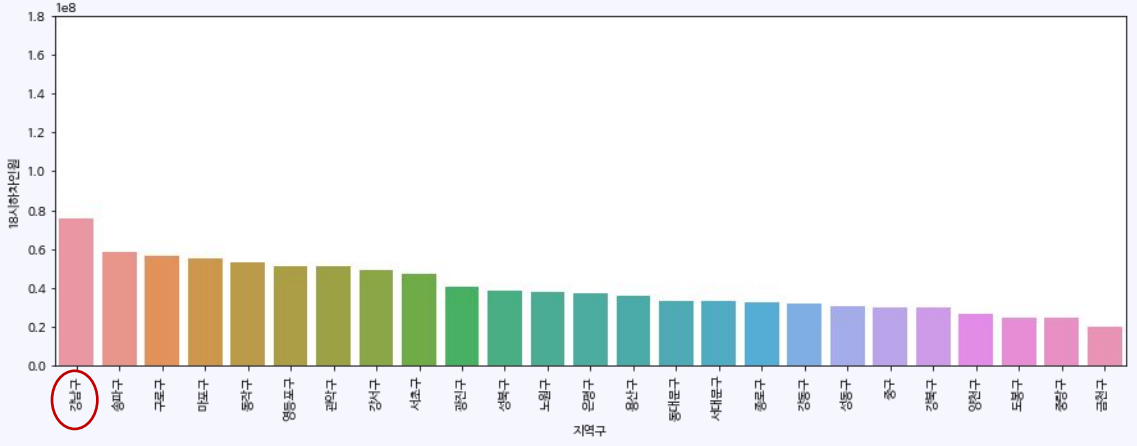
승차





하차





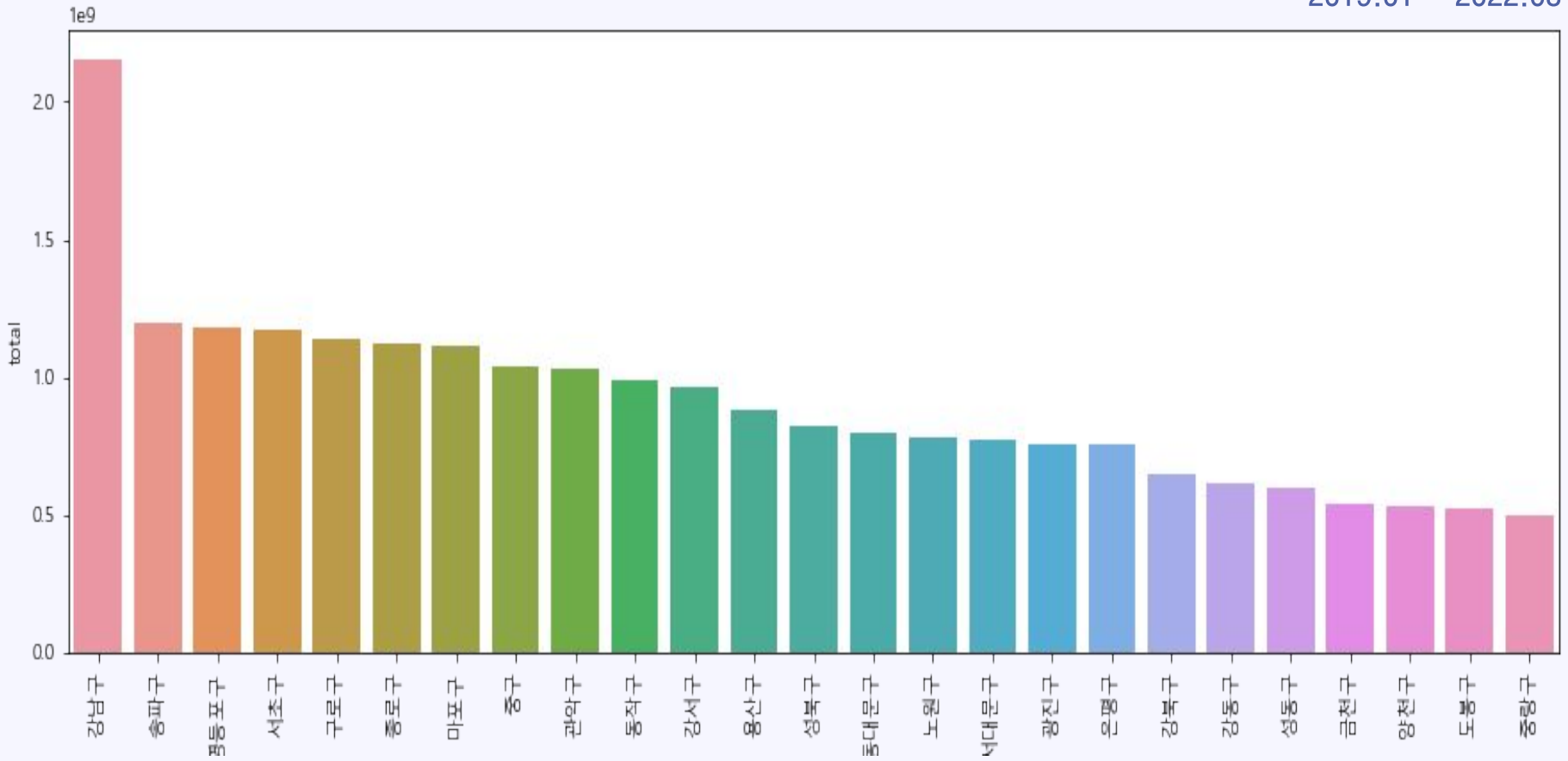
승차 > 구로구
하차 > 강남구

승차 > 강남구
하차 > 강남구

Data Charts Infographics (Region)

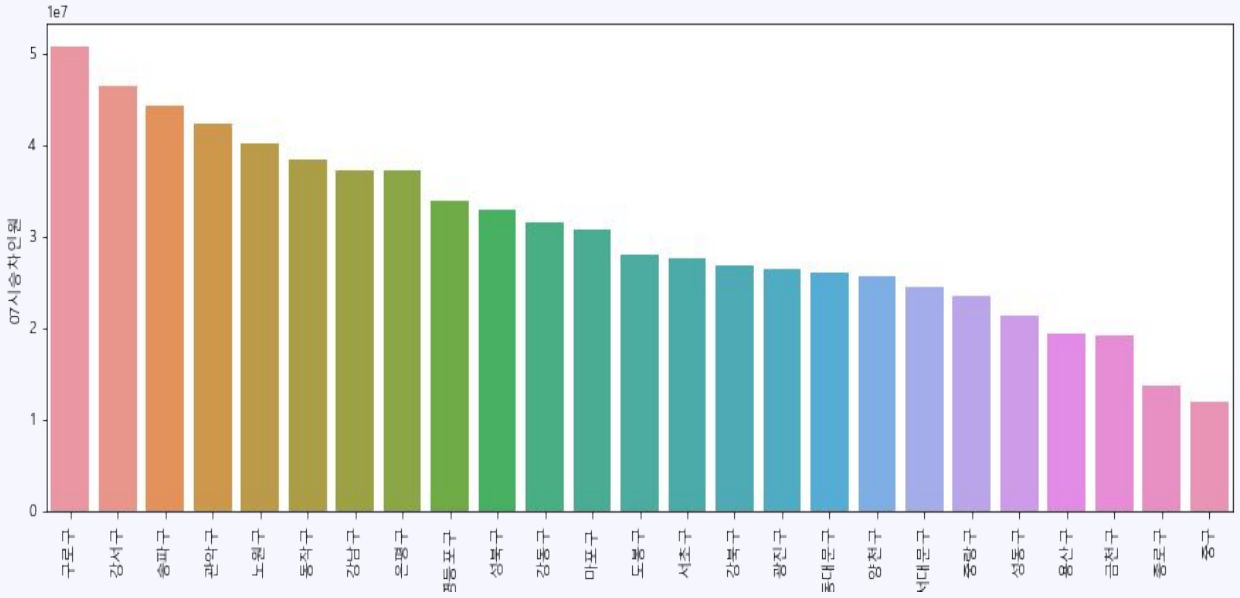
지역별 승/하차 총 인원

2019.01 ~ 2022.08

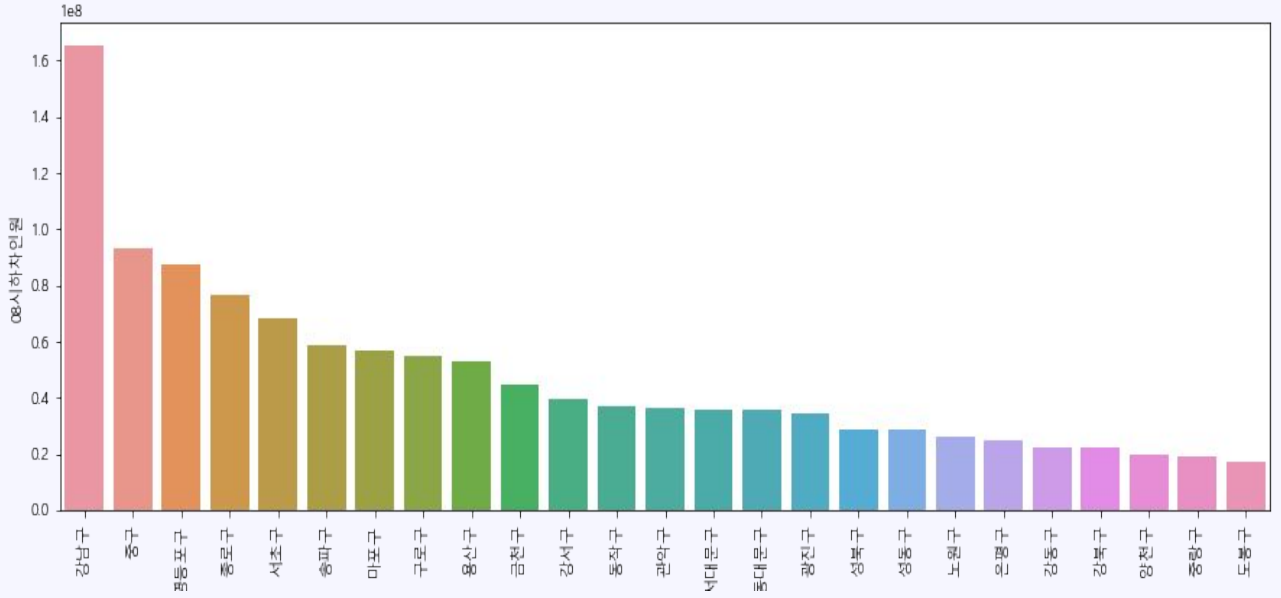


Data Charts Infographics (출근시간)

2019.01 ~
2022.08



07시 승차인원

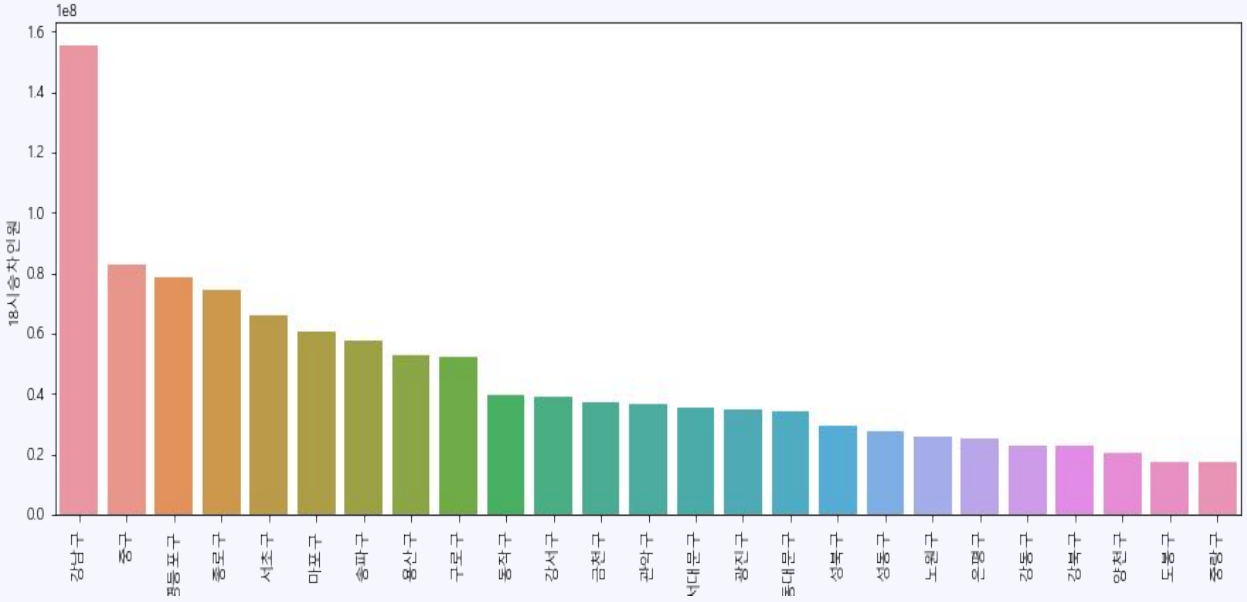


08시 하차인원

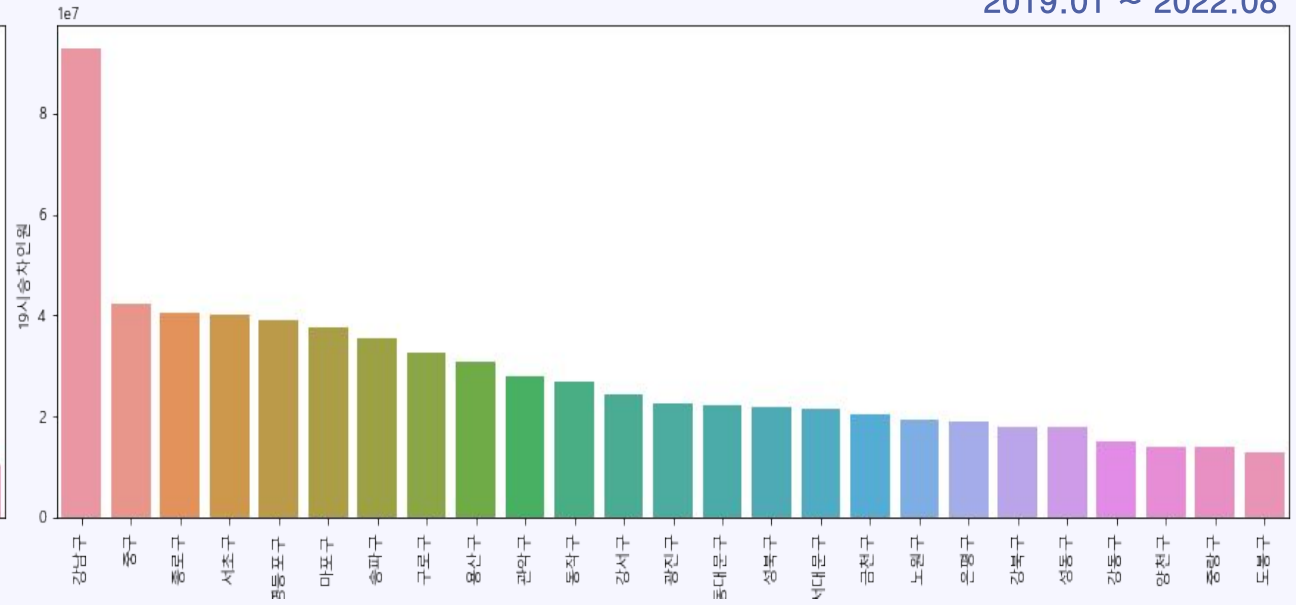
>> **구로구**에서 승차인원이 제일 많고,
강남구에서 하차인원이 제일 많다.

Data Charts Infographics (퇴근시간)

2019.01 ~ 2022.08



18시 승차인원

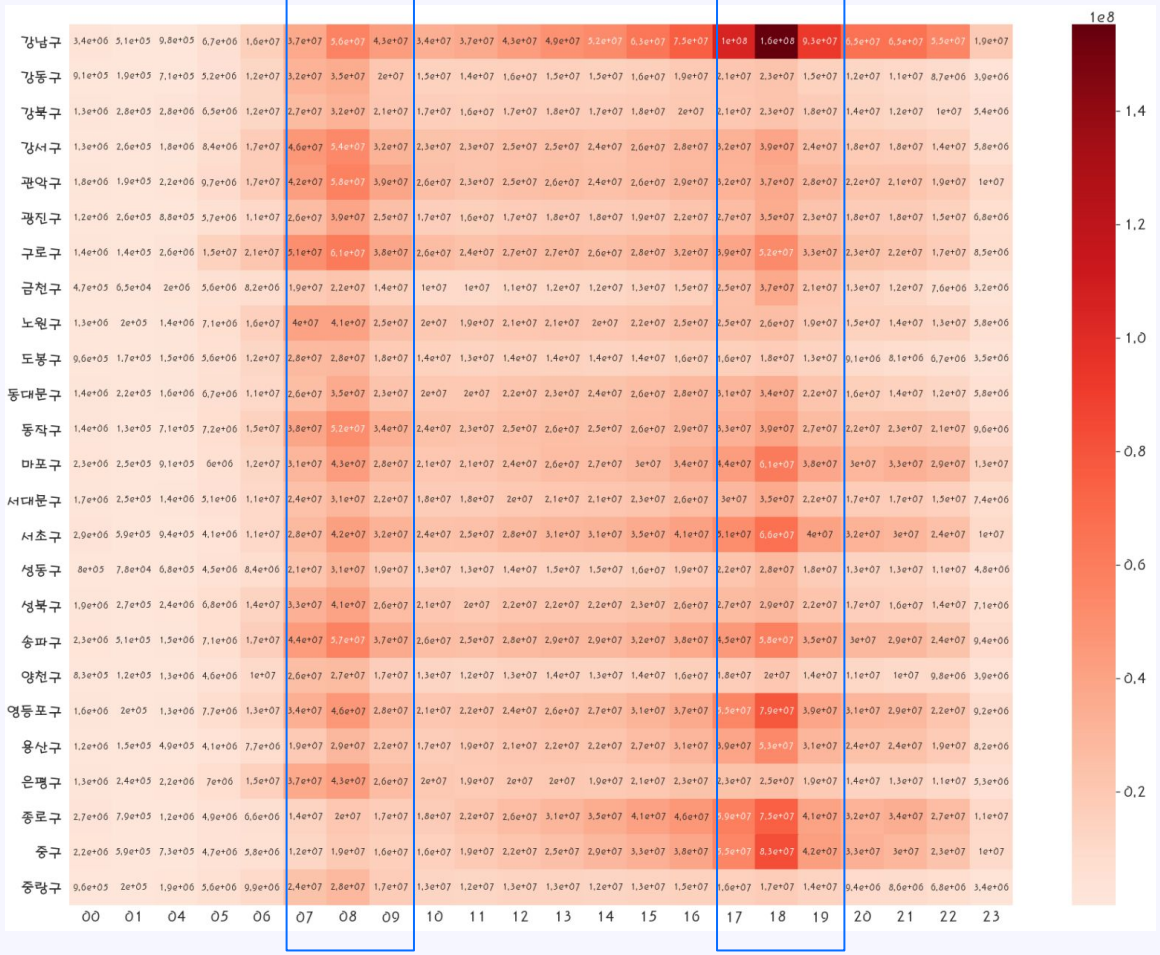


19시 승차인원

>> **강남구**에서 승차인원이 제일 많다.

Data Charts Infographics (Region)

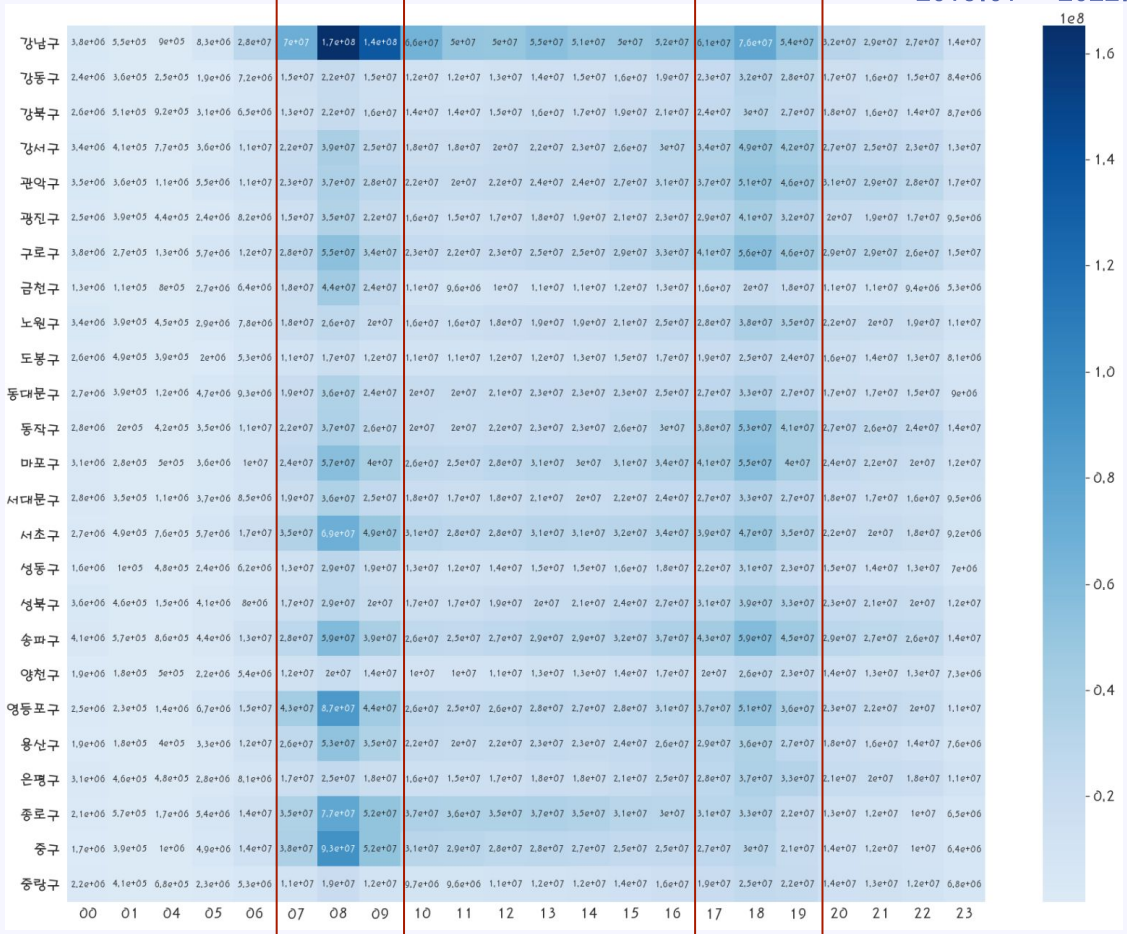
승차인원



출근시간 퇴근시간

하차인원

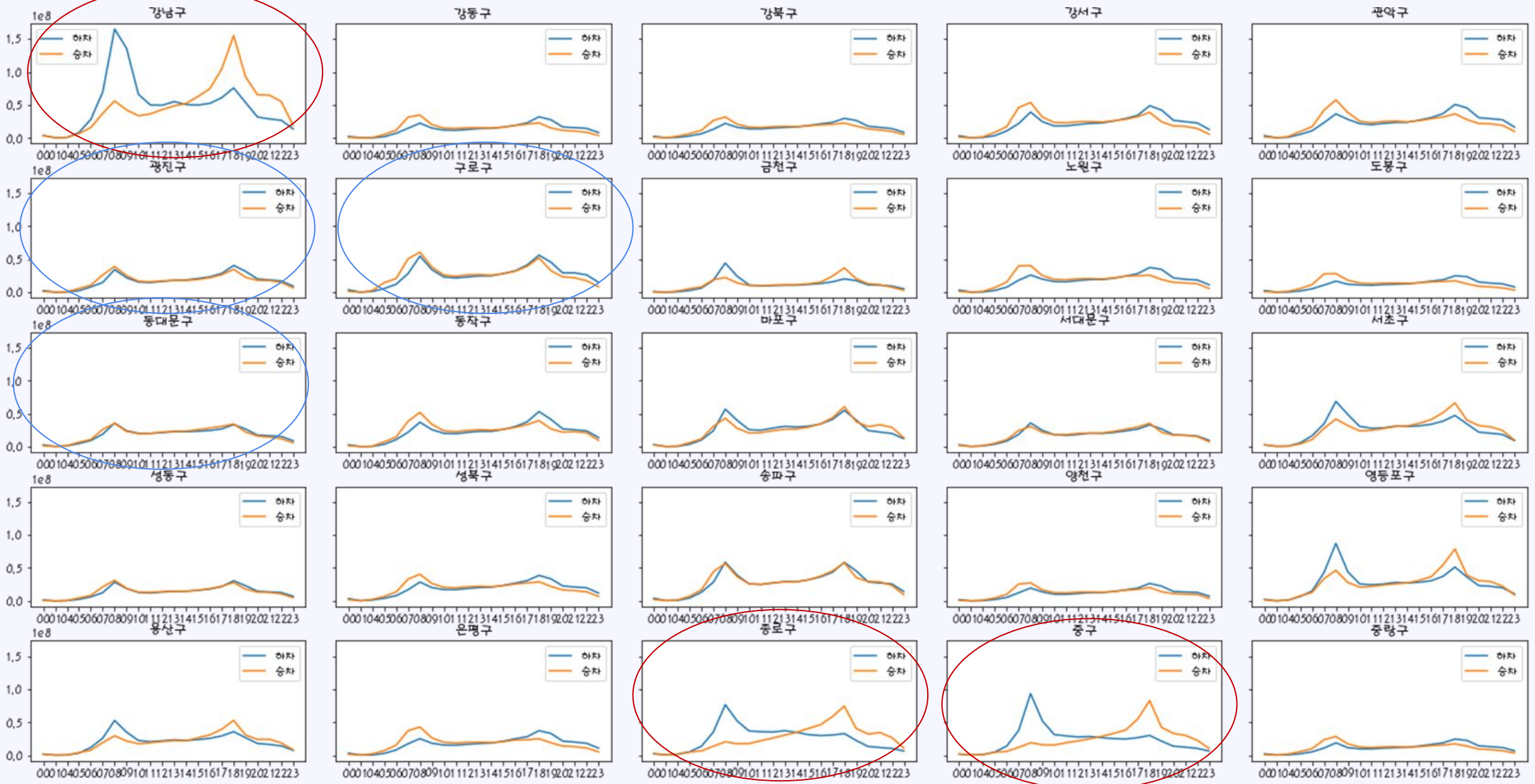
2019.01 ~ 2022.08



출근시간 퇴근시간

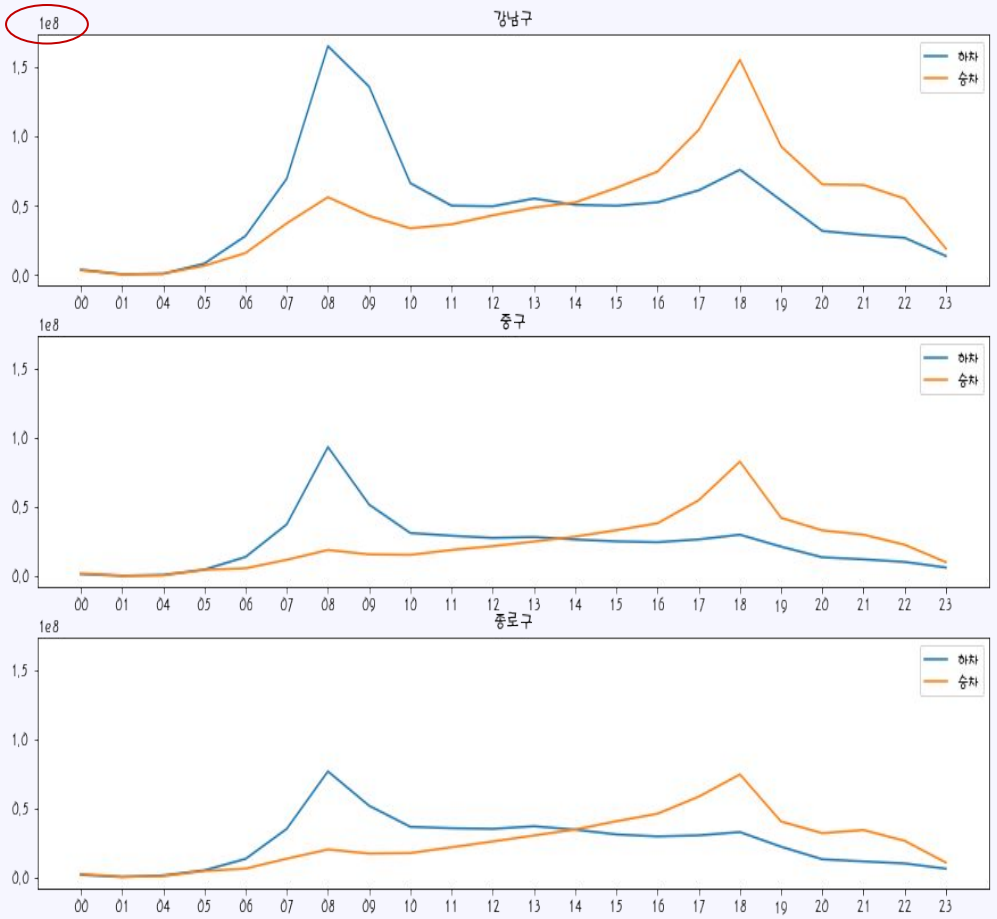
Data Charts Infographics (Region)

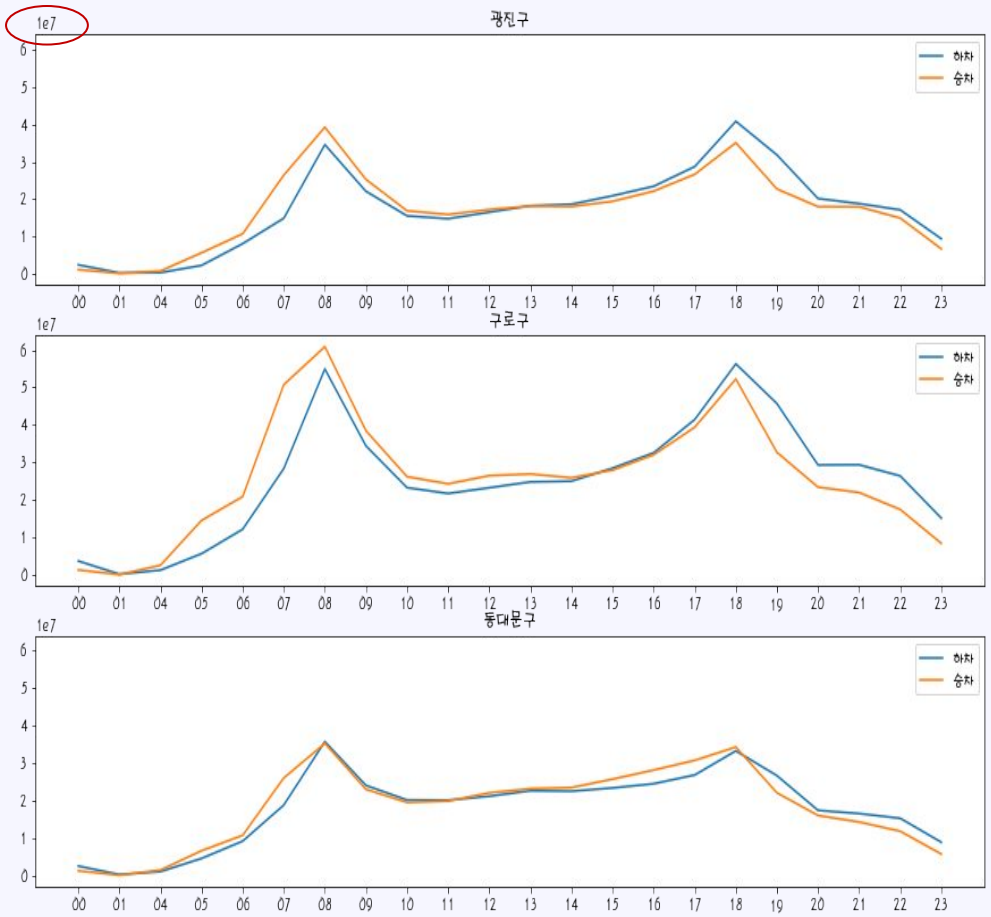
2019.01 ~ 2022.08



Data Charts Infographics (Region)

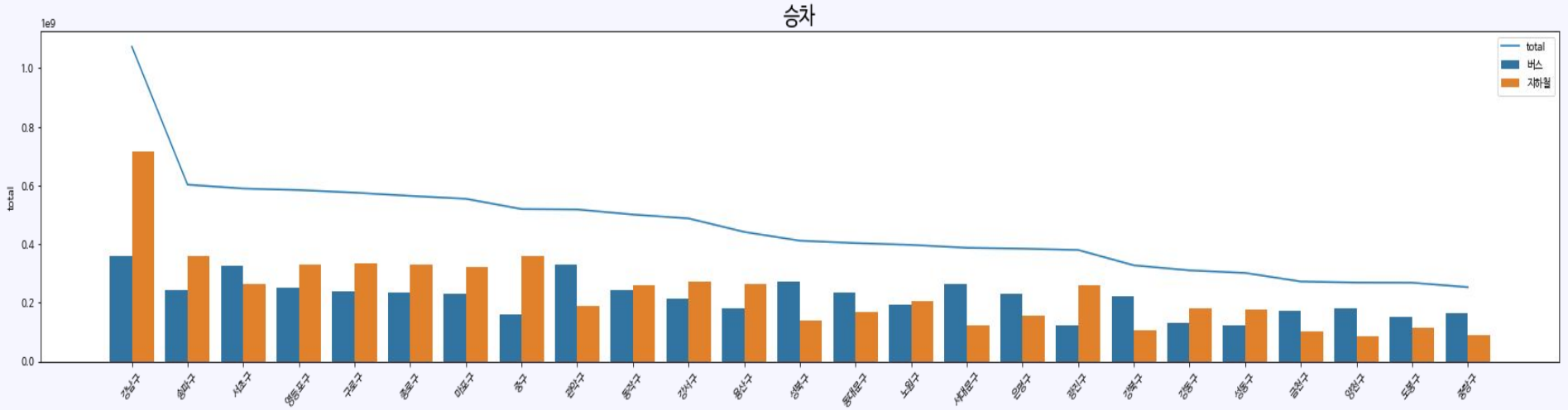
2019.01 ~ 2022.08

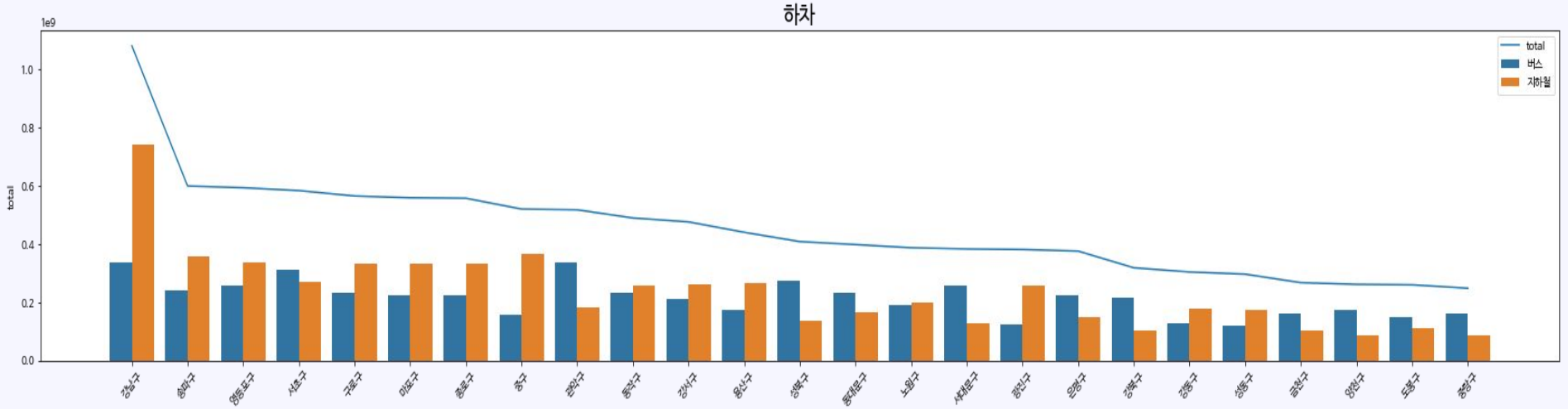




- **강남구, 중구, 종로구(회사)**에서 승차,하차 차이가 제일 많고,
- **광진구, 구로구, 동대문구(주거)**에서 승차,하차 차이가 제일 적다.

Data Charts Infographics (Region)





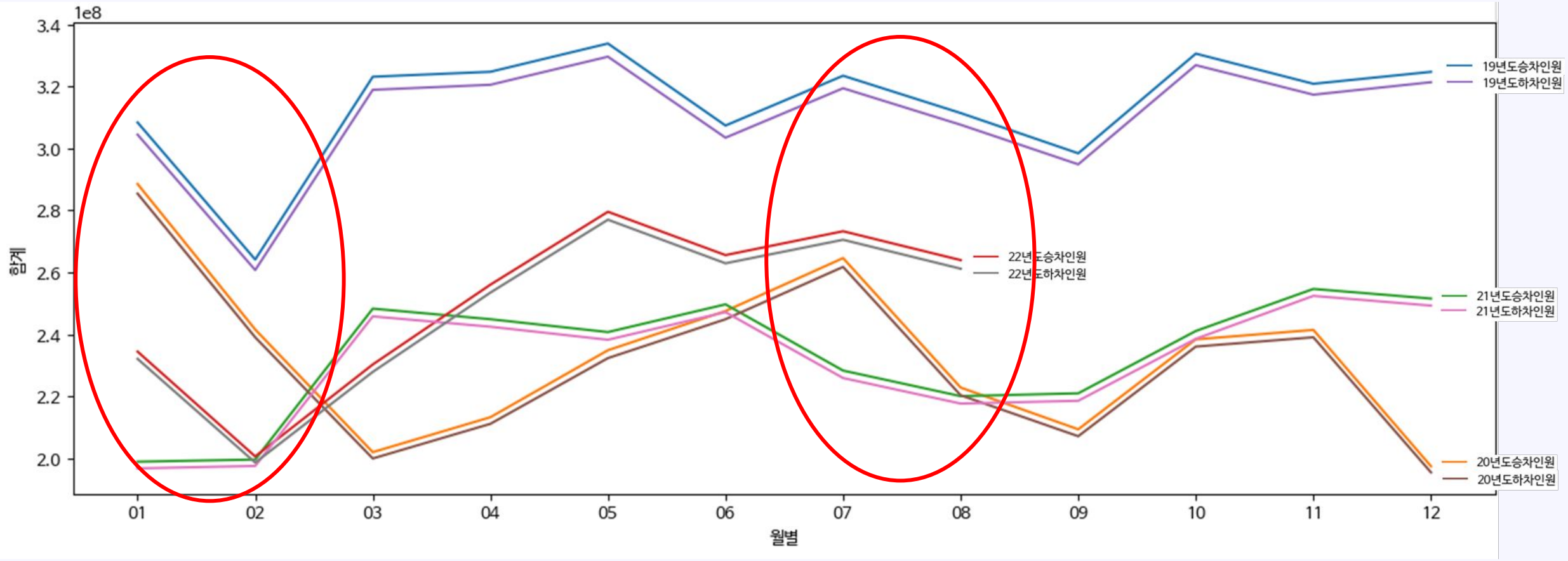
1. 지하철역 대비 버스정류장 수 많으나 이용 인원은 지하철이 많다.

2. 강남구에서 승하,하차 인원이 다른지역보다 월등히 높은 이유는 지하철

3. 지역구별 버스이용객 차이는 많이 나지 않는다.

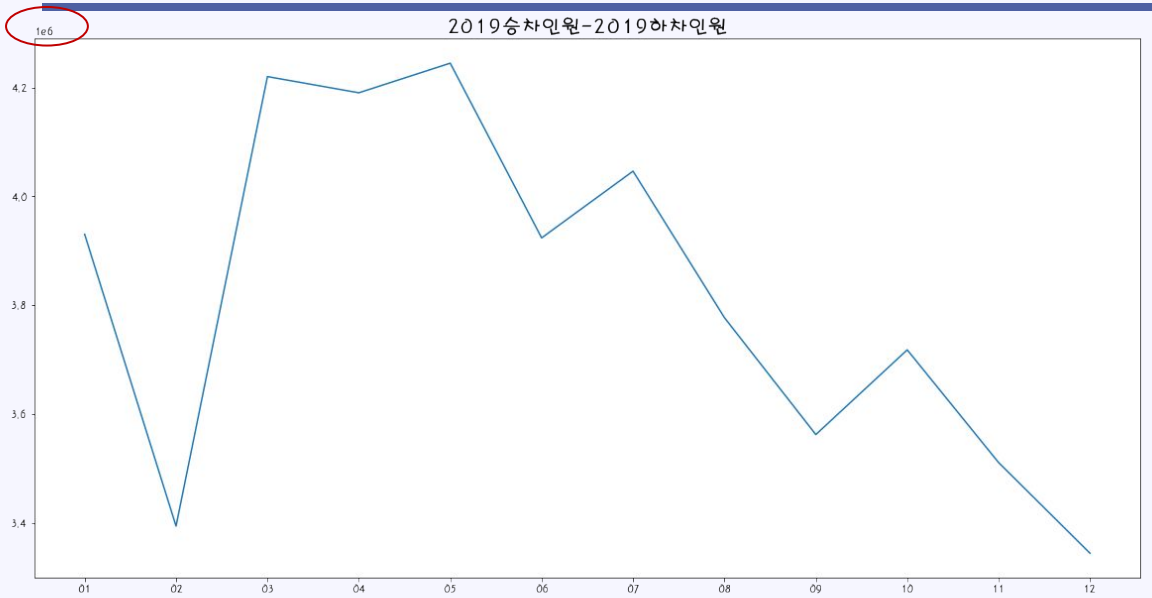
Data Charts Infographics (Period)

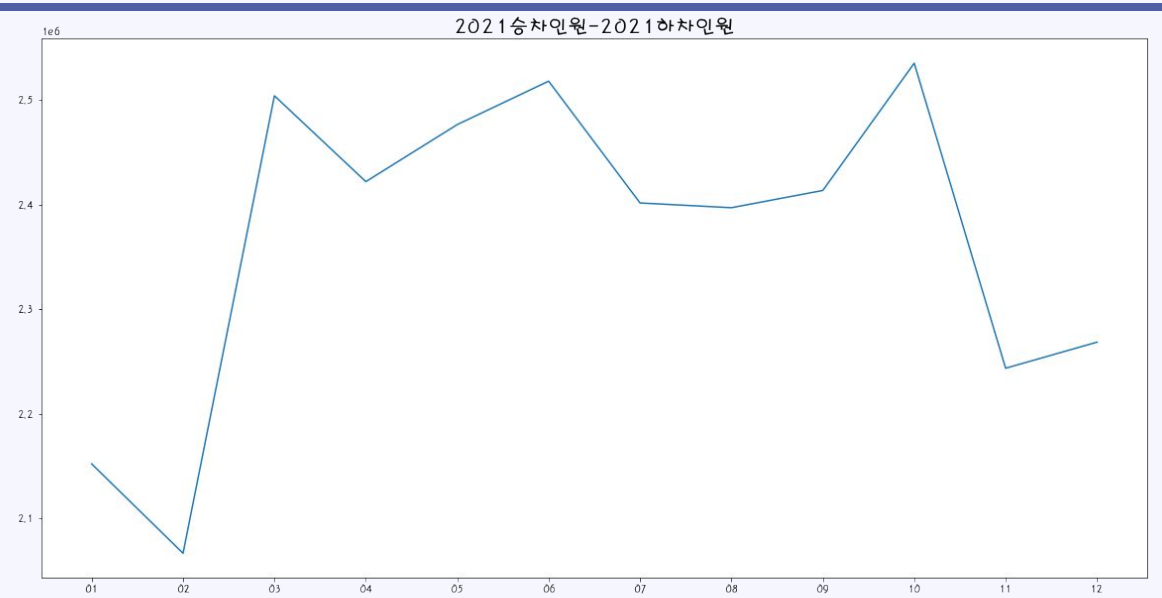
월별 승/하차 총 인원

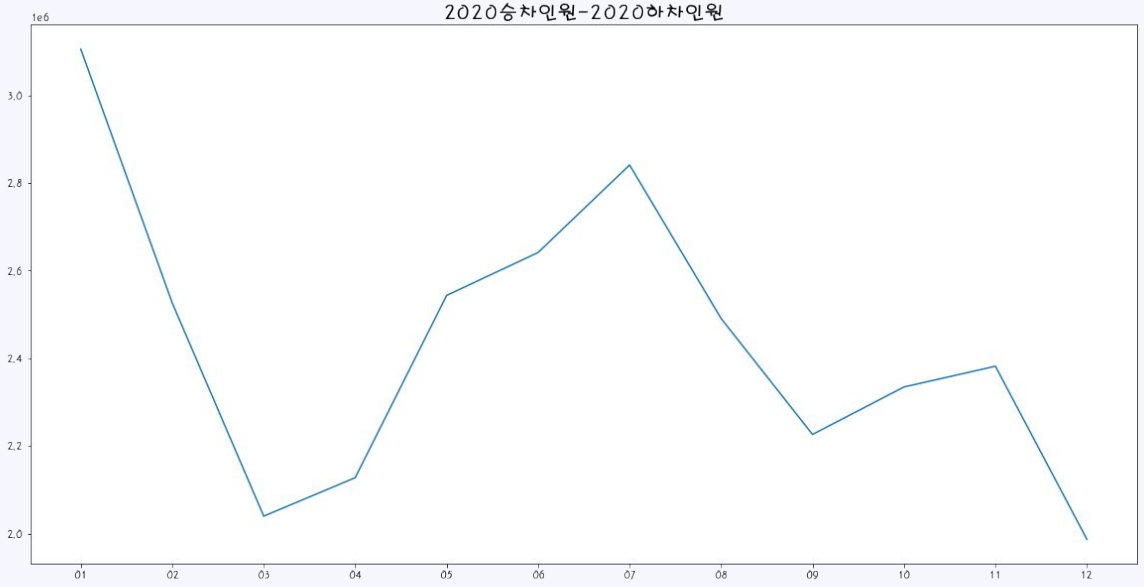


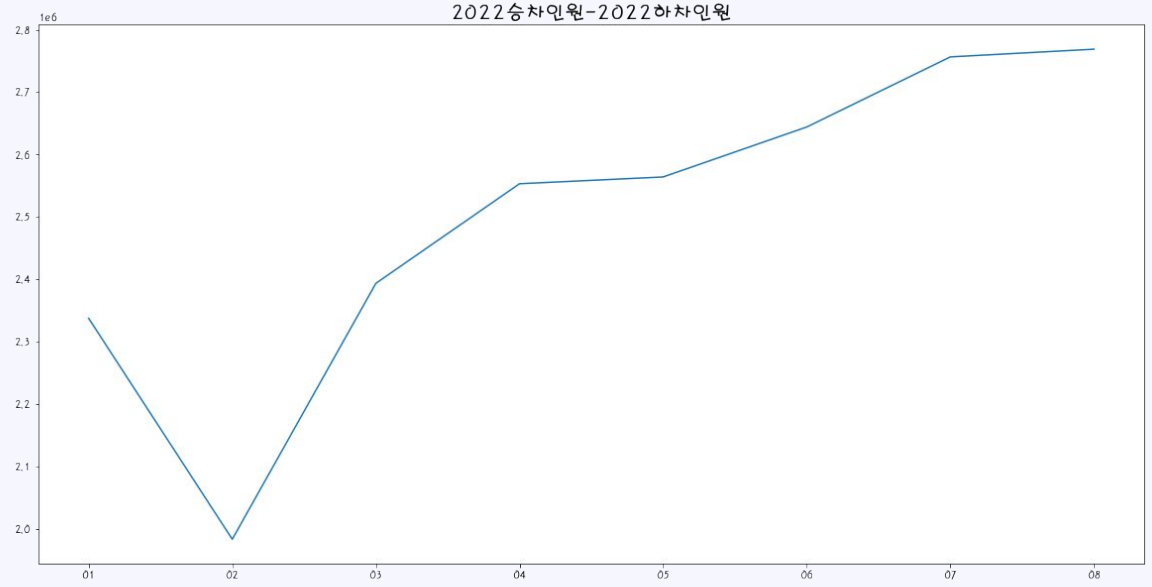
1,2,7,8월>방학(학교) 비교적 이용 저하

Data Charts Infographics (Period)



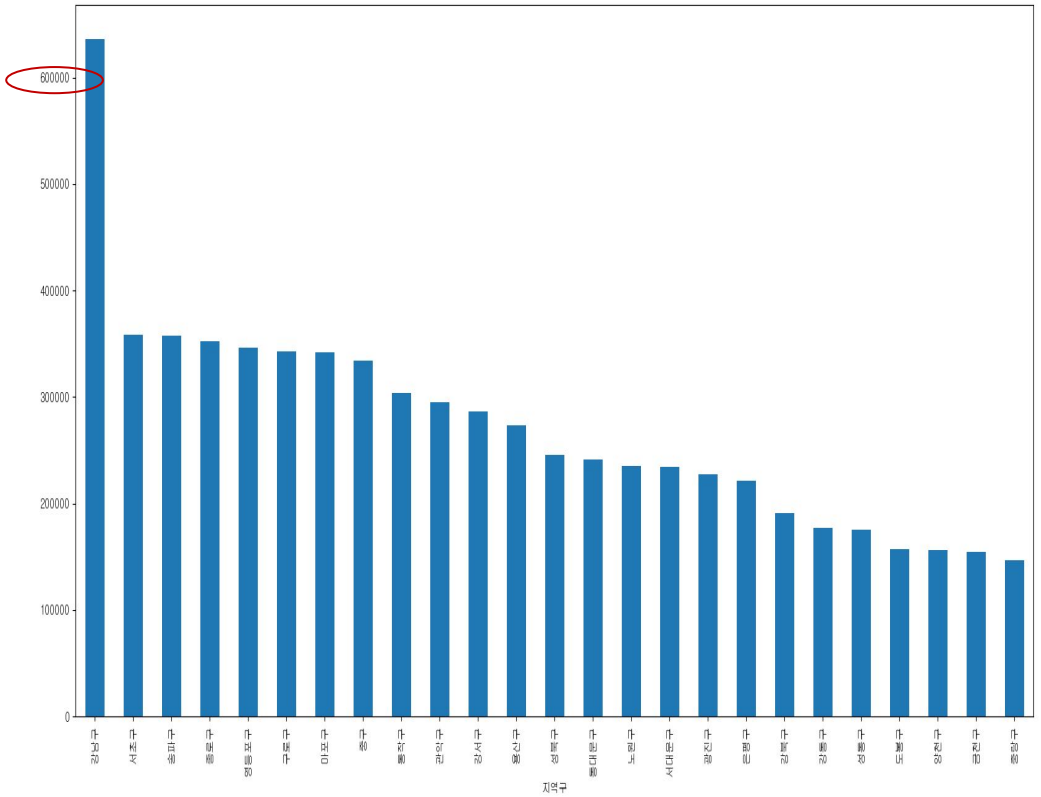




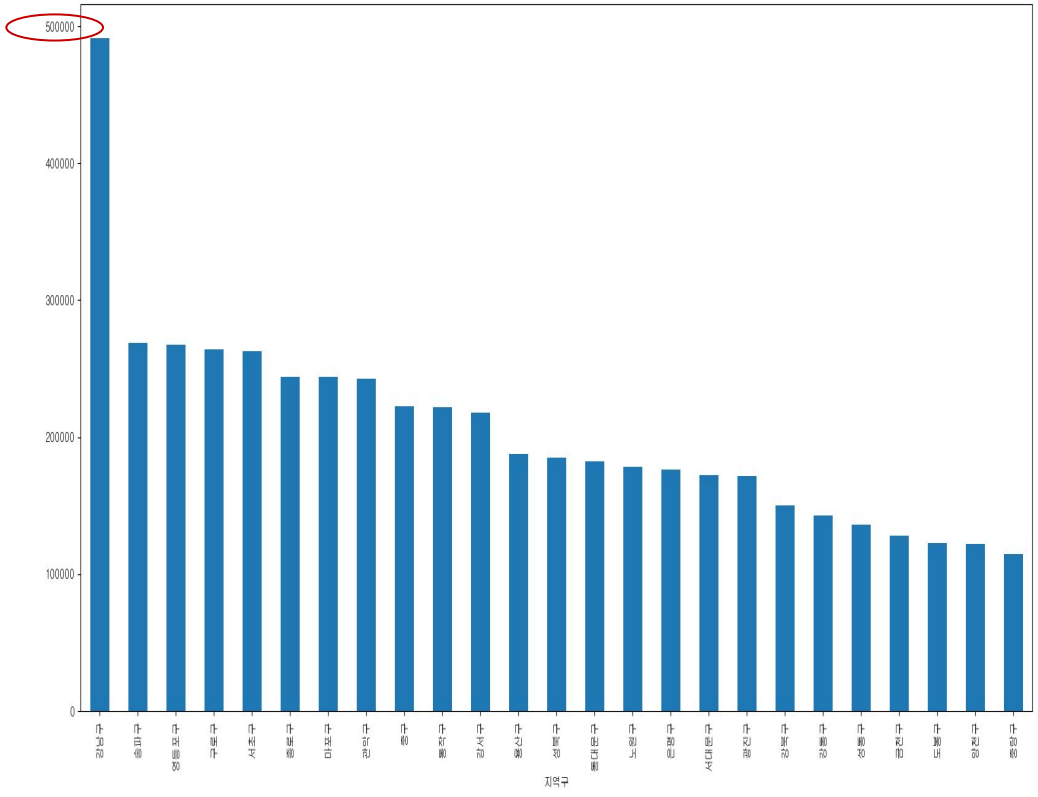


Data Charts Infographics (Period)

코로나 강화정책 전후 변화



2019.03~2020.03



2020.04~2021.04

Conclusion

- 출근시간에는 08시 하차인원이 제일 많고 퇴근시간에는 18시 승차인원이 제일 많다.
- 출근시간이 사람마다 다 다르지만 07시,08시 두 시간대 모두 구로구에서 승차인원이 제일 많다.
- 퇴근시간에 18시 승차,하차 인원 모두 강남에서 제일 많다.

  >> 2019~2022 지역별 총 승하차인원 > 강남에서 제일 많다.
- 계절(월별)에 따른 변화 존재(1,2,7,8월은 비교적 이용 저하)
- 시간에 따른 승하차 차이가 많은 강남구,중구,종로구는 회사들이 밀집되어 있을 가능성이 존재 有/

  차이가 많이 나지 않은 광진구,구로구,동대문구는 주거지역이 밀집되어 있을 확률이 크다라고 추정.

  ex) 거주지역 → 08시 출근후 20시 귀가

Discussion (Limitations & Research Directions)

Limitations

- ✔ 환승역의 밀집도는 확인 불가
- ✔ 경기도와 그 외의 지역까지 확인 불가

Research Directions

- ✔ 코로나 이전과 이후, 각 거리두기 변화에 따른 이용량의 차이
- ✔ 회사 수, 상업 단지 등이 대중교통 이용량에 미치는 영향