

빅 데이터 분석 결과 시각화

지하철 무임 승차현황 분석

이
조
최

지
예
희

영
슬
경

CONTENTS

01

데이터 이해

- 데이터 선정
- 데이터 수집
- 데이터관련 현황

02

분석 전 처리

- 분석 전 가설
- 파일 전 처리
- 변수 추가 확정
- 위치정보 전처리

03

EDA 및 시각화

- 추세확인
- 분포확인
- 상관관계
- 코로나영향 확인

04

INSIGHT

- 시각화를 통한 통찰
- Reflection

01

데이터 선정 방향성

- 이해하기 쉽고 사회적 이슈가 있는 데이터 선정
- 웹 스크래핑(web scraping) 보다는 제공된 공공API등을 이용
- 시대현황을 반영할 수 있는 데이터

데이터 수집

- 서울시 열린 데이터 광장 (<http://data.seoul.go.kr>)
- 서울시 지하철 호선별 역별 유무임 승하차 인원정보.csv
- 서울시 역코드로 지하철역 위치 조회.csv

데이터 현황

- 서울시 지하철 호선별 역별 유무임 승하차 인원정보.csv
 - ✓ 43687 건의 데이터가 8개의 변수로 이루어져 있음
 - ✓ 2015년1월 ~ 2020년4월까지의 데이터
- 서울시 역코드로 지하철역 위치조회.csv
 - ✓ 929개의 전철역코드와 위도경도 정보

02

분석 전 가설

- 2015년 부터 현재까지의 시계열 변화 추이를 확인하여 추세예측
- 2020년 2월~ 4월 자료를 분석하여 코로나 이후의 변화 추정
- 이용객이 많은 혼잡역사의 데이터 분석 후 대응방안 제안
- 무임 승하차가 많은 역사의현황을 파악하고 역사 별 대응 계획 제안

데이터 현황(탐색)

	사용월	호선명	지하철역	유임승차인원	무임승차인원	유임하차인원	무임하차인원	작업일자
0	202004	9호선2~3단계	언주	184170	16729	185046	16460	20200503
1	202004	2호선	합정	680012	48502	732794	48049	20200503
2	202004	2호선	을지로3가	399344	51720	403589	50877	20200503
3	202004	2호선	강변(동서울터미널)	676082	83345	673317	81844	20200503

	사용월	호선명	지하철역	유임승차인원	무임승차인원	유임하차인원	무임하차인원	작업일자
43685	201501	6호선	증산	244352	226314	65797	64899	20150206
43686	201501	4호선	숙대입구	426269	421753	59564	56862	20150206

02

데이터 전처리

```
subway = pd.read_csv('subway.csv', encoding='CP949')
```

작업일자 제외

```
subway = subway.iloc[:,0:7]
```

중복데이터 제거

```
subway = subway.drop_duplicates(['사용월', '호선명', '지하철역', '유임승차인원', '무임승차인원', '유임하차인원', '무임하차인원'])
```

연도, 월 분리

```
subway["연도"] = subway["사용월"].astype(str).str[:4]
subway["월"] = subway["사용월"].astype(str).str[4:]
```

9호선2~3단계, 9호선2단계 >> 9호선 으로 변경

```
subway.loc[subway['호선명'] == '9호선2~3단계', '호선명'] = '9호선'
subway.loc[subway['호선명'] == '9호선2단계', '호선명'] = '9호선'
```

1호선~9호선 데이터만 선택

```
sub1 = subway[subway['호선명'] == '1호선']
sub2 = subway[subway['호선명'] == '2호선']
sub3 = subway[subway['호선명'] == '3호선']
sub4 = subway[subway['호선명'] == '4호선']
sub5 = subway[subway['호선명'] == '5호선']
sub6 = subway[subway['호선명'] == '6호선']
sub7 = subway[subway['호선명'] == '7호선']
sub8 = subway[subway['호선명'] == '8호선']
sub9 = subway[subway['호선명'] == '9호선']

sub = pd.concat([sub1, sub2, sub3, sub4, sub5, sub6, sub7, sub8, sub9]).reset_index(drop=True)
```

만 명 단위로 변경

```
col_list = ['유임승차인원', '무임승차인원', '유임하차인원', '무임하차인원']
sub[col_list] = sub[col_list]/10000
```

필요 없는 열제거

중복된 행 제거(시각화 중 발견)

사용년월 → 연도 와 월로 분리
타입변경(category)

9호선 개발단계별 다른 이름 통일

1~ 9호선 데이터 활용
(경강선,중앙선 등 제외)

단위 조정(만명)

02

지도 시각화 – 위도경도 데이터 전처리

	X좌표	X좌표(WGS)	Y좌표	Y좌표(WGS)	사이버스테이션	외부코드	전철역명	호선
0	525992.0	37.492522	1108579.0	127.118234	2818.0	817	가락시장	8
1	525992.0	37.492522	1108579.0	127.118234	2818.0	350	가락시장	3
2	498060.0	37.571607	1130332.0	126.991806	153.0	534	종로3가	5
3	498060.0	37.571607	1130332.0	126.991806	153.0	329	종로3가	3
4	498060.0	37.571607	1130332.0	126.991806	153.0	130	종로3가	1

지하철역 위도, 경도 정보 파일 확인

	X좌표(WGS)	Y좌표(WGS)
전철역명		
419민주요지	NaN	NaN
가남	37.748577	127.044213
가락시장	37.492522	127.118234
가산디지털단지	37.481072	126.882343
가양	37.561391	126.854456

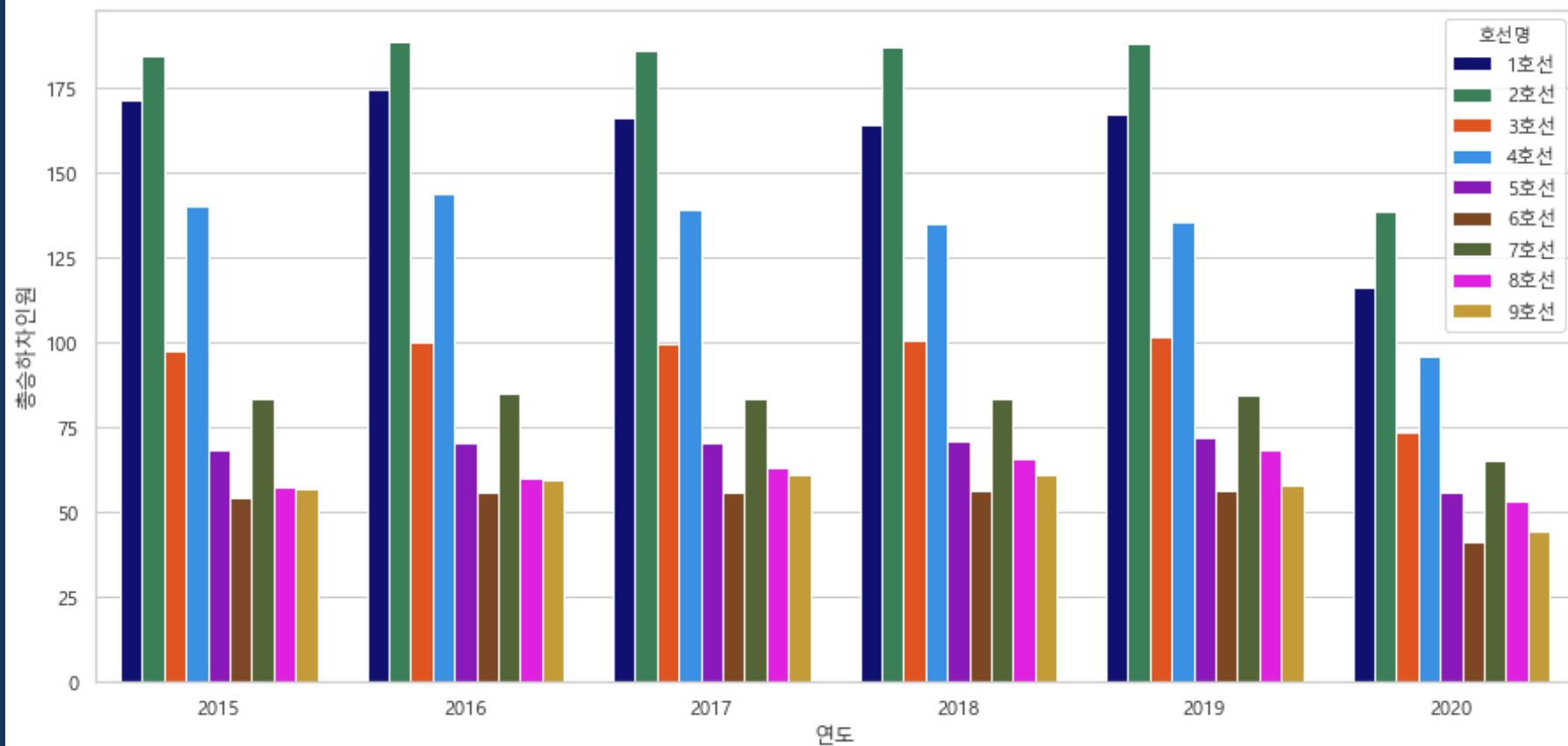
전철역명으로 groupby 위도, 경도 데이터만 취함

	유임승하차인원	무임승하차인원	승하차인원	X좌표(WGS)	Y좌표(WGS)	무임비율
count	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000
mean	2704.783275	491.310646	3196.093921	37.524563	126.989450	15.762198
std	861.023953	112.242411	938.186620	0.043932	0.065203	2.935255
min	1799.432288	295.356538	2211.987308	37.476530	126.882343	10.788215
25%	2150.477830	419.250848	2560.047577	37.485000	126.935099	13.301616
50%	2498.266276	464.023577	2956.585147	37.508784	126.984624	16.247867
75%	3021.978125	562.980083	3539.081478	37.558141	127.035087	17.238538
max	5323.427051	719.862615	6029.979808	37.638052	127.102234	23.176028

분석용 데이터와 결합

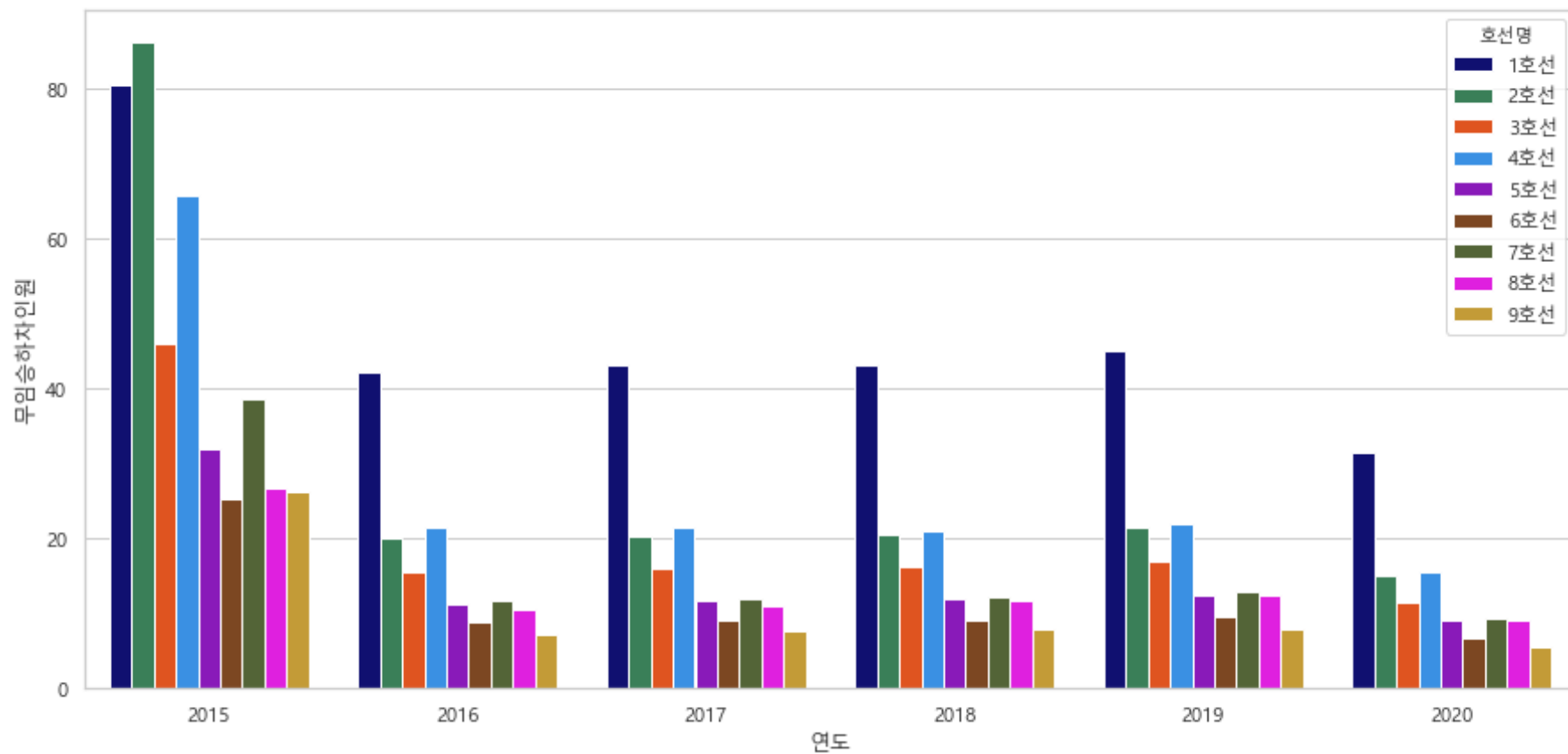
03

연도별 호선별 총 승하차 인원 (bar plot)



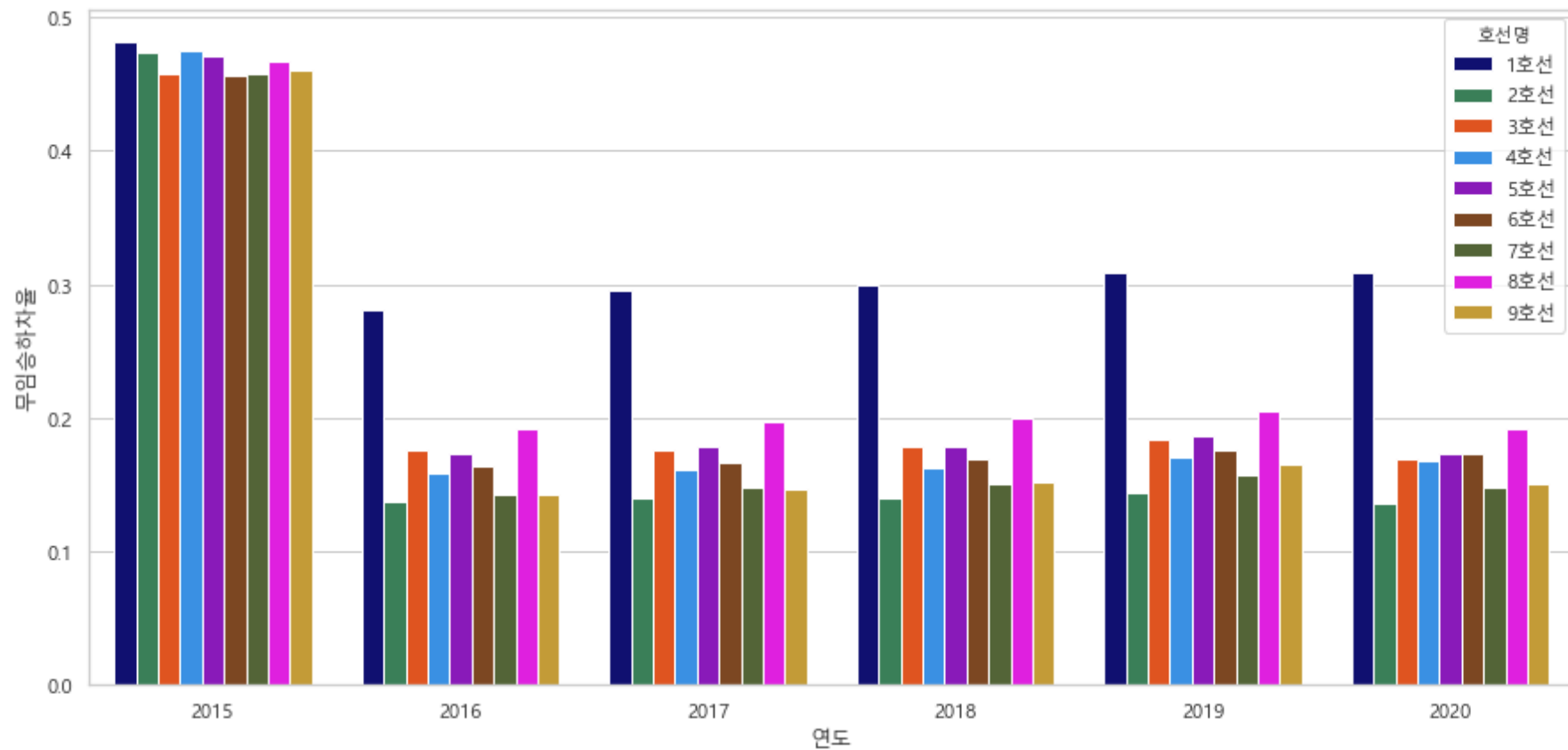
03

연도별 호선별 무임 승하차 인원



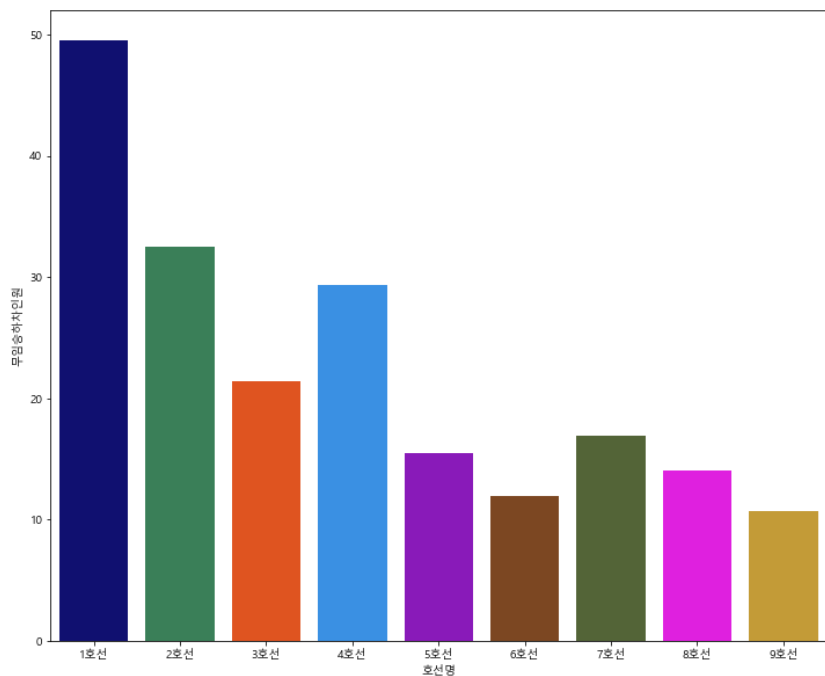
03

연도별 호선별 무임승하차율

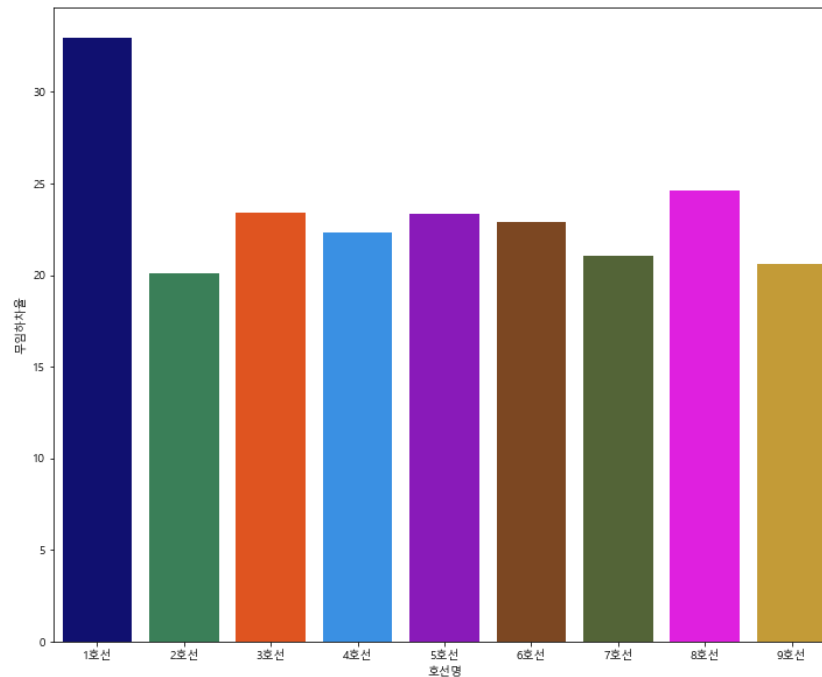


03

호선별 무임승하차인원



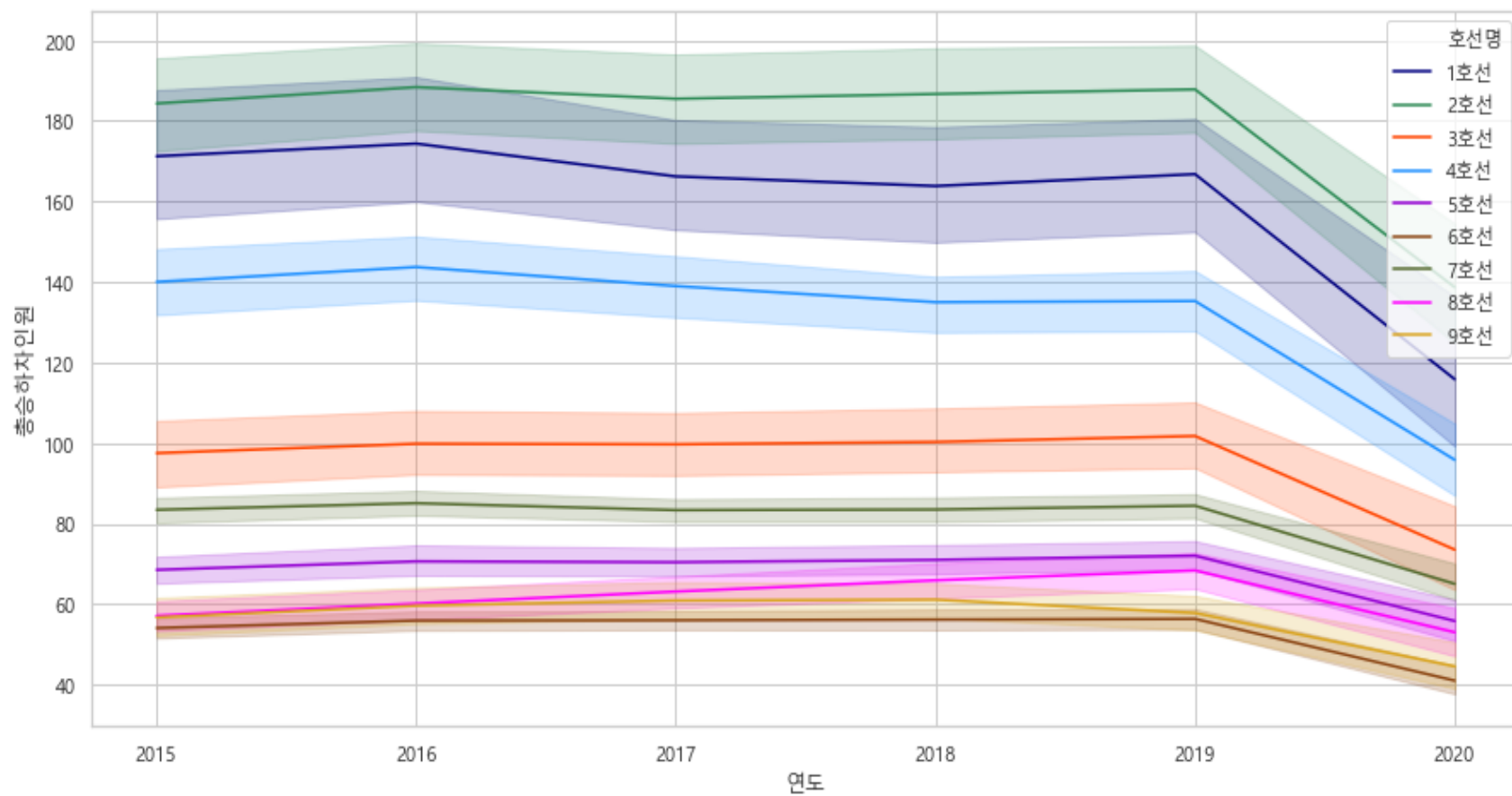
호선별 무임하차율



기간 : 2015년부터 2020년 4월까지

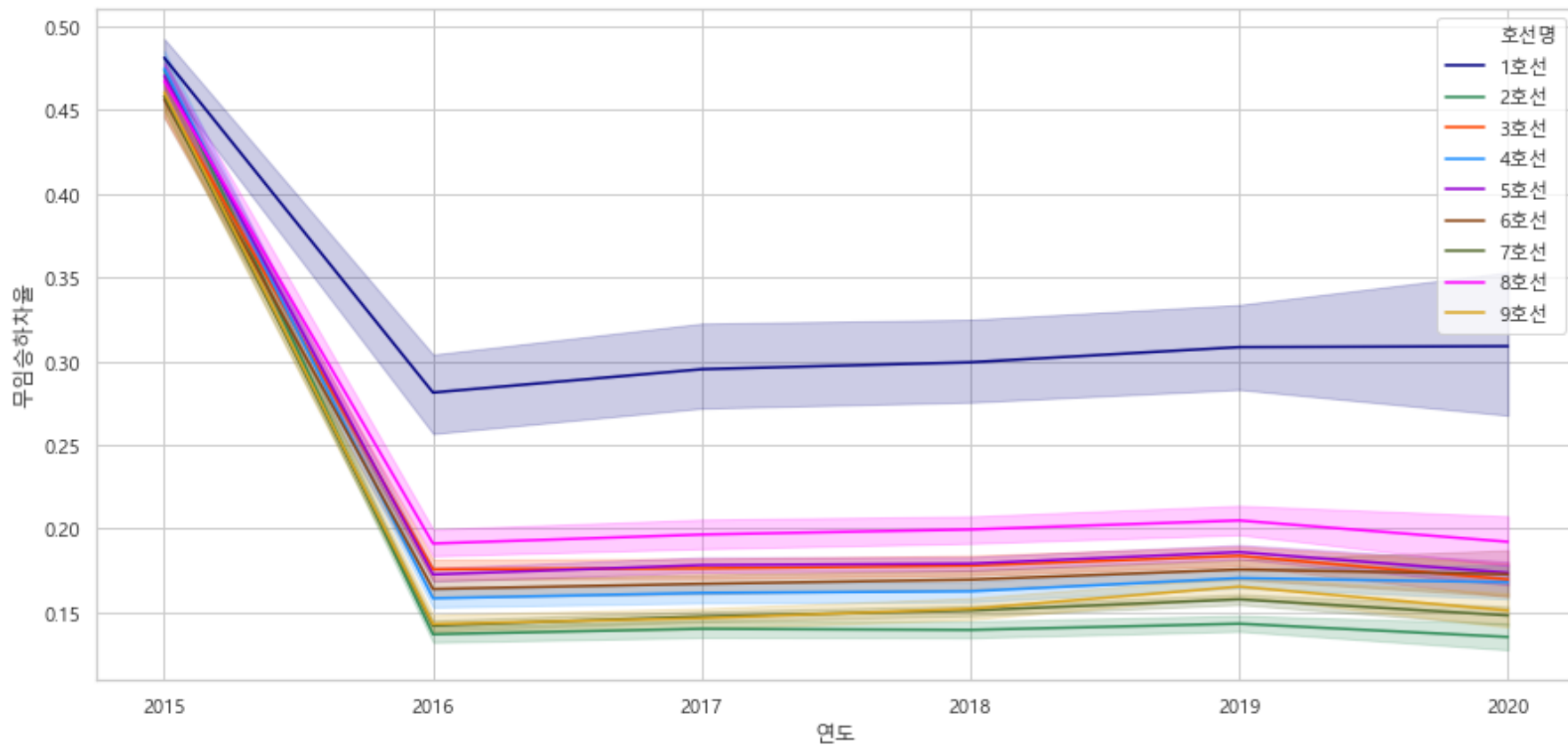
03

연도별 호선별 승하차 인원 (line plot)



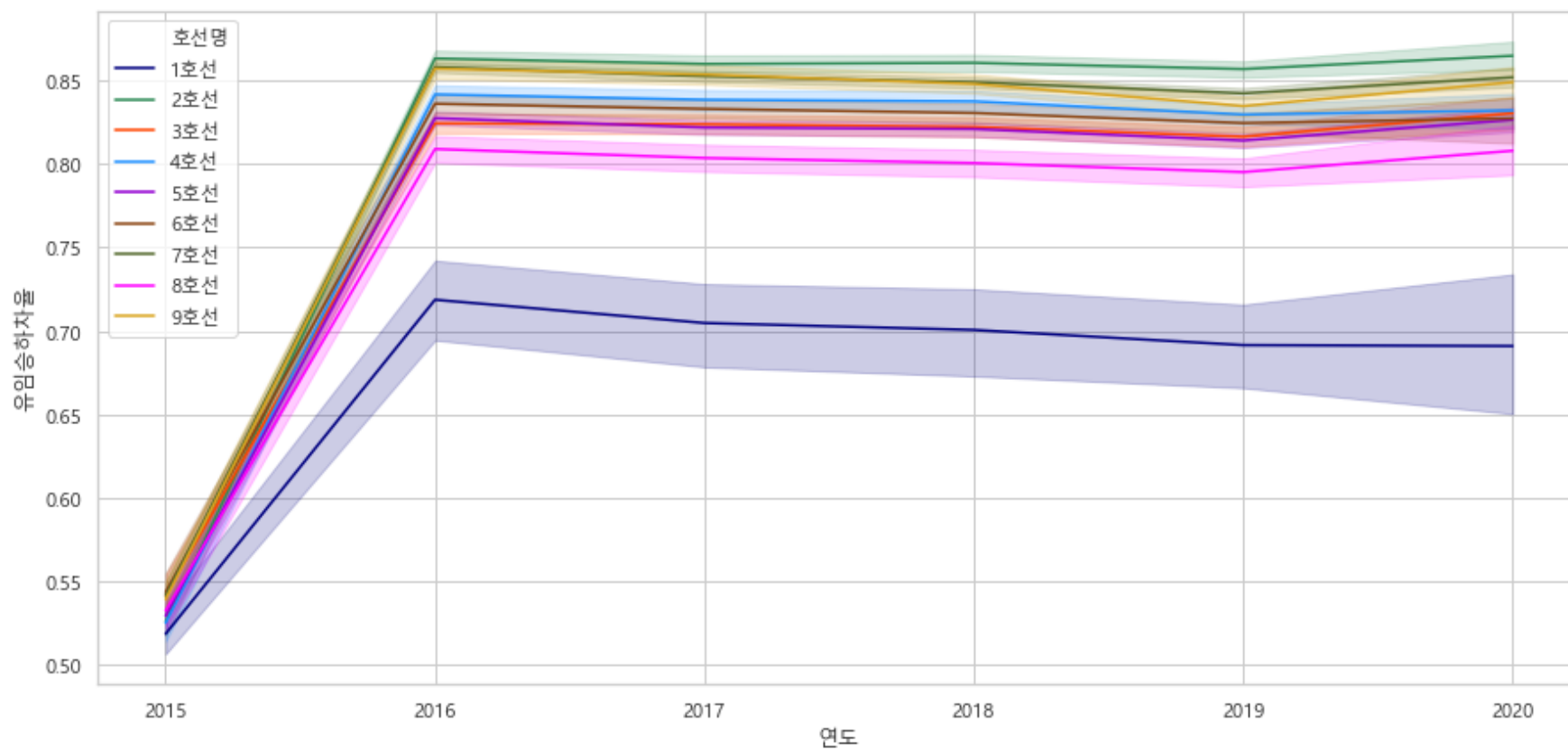
03

연도별 호선별 무임승하차율 (line plot)



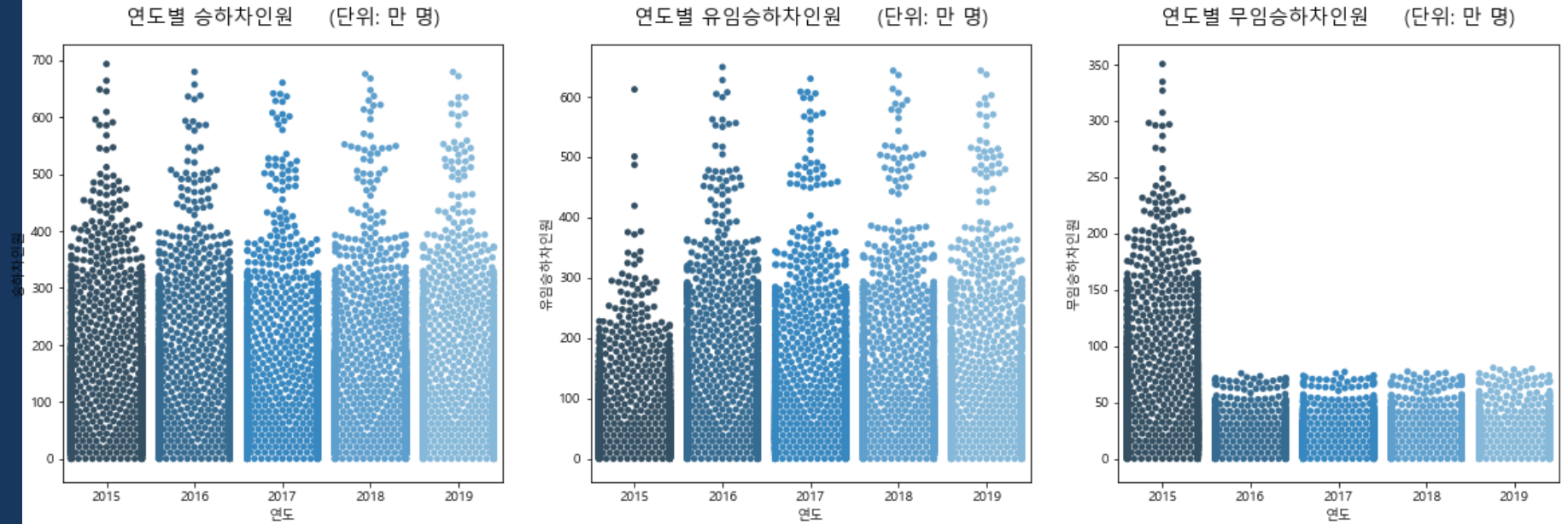
03

연도별 호선별 유임승하차율 (line plot)



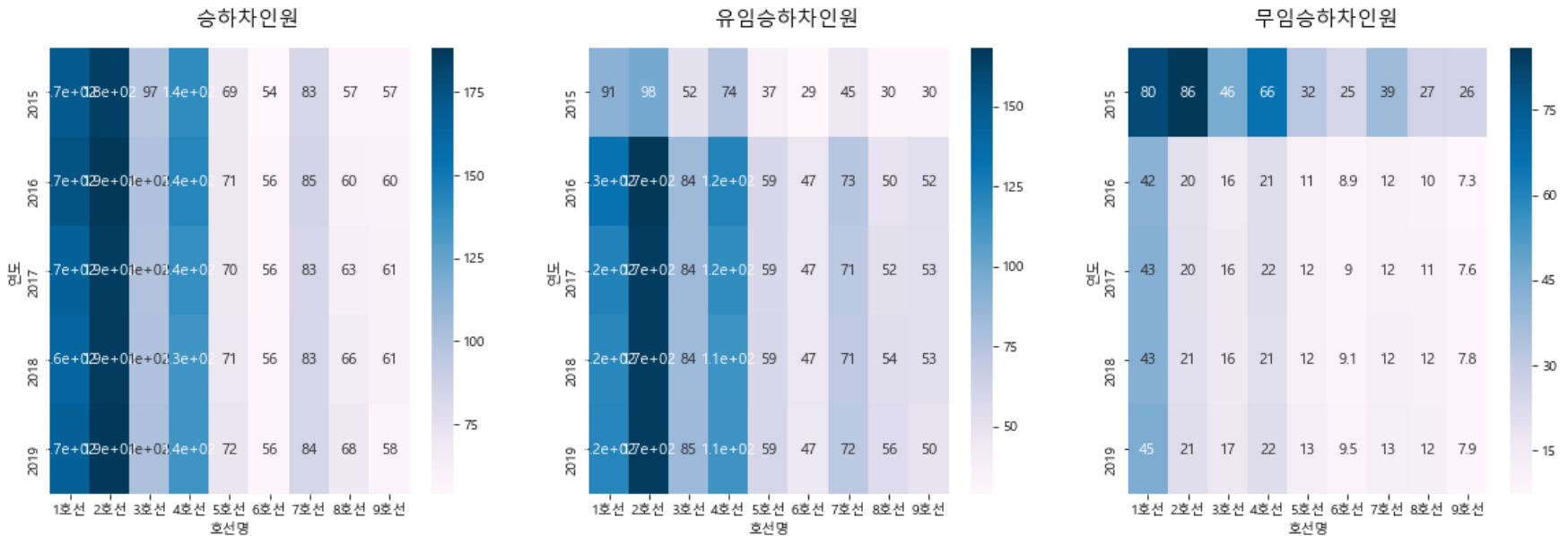
03

연도별 승하차인원의 swamplot

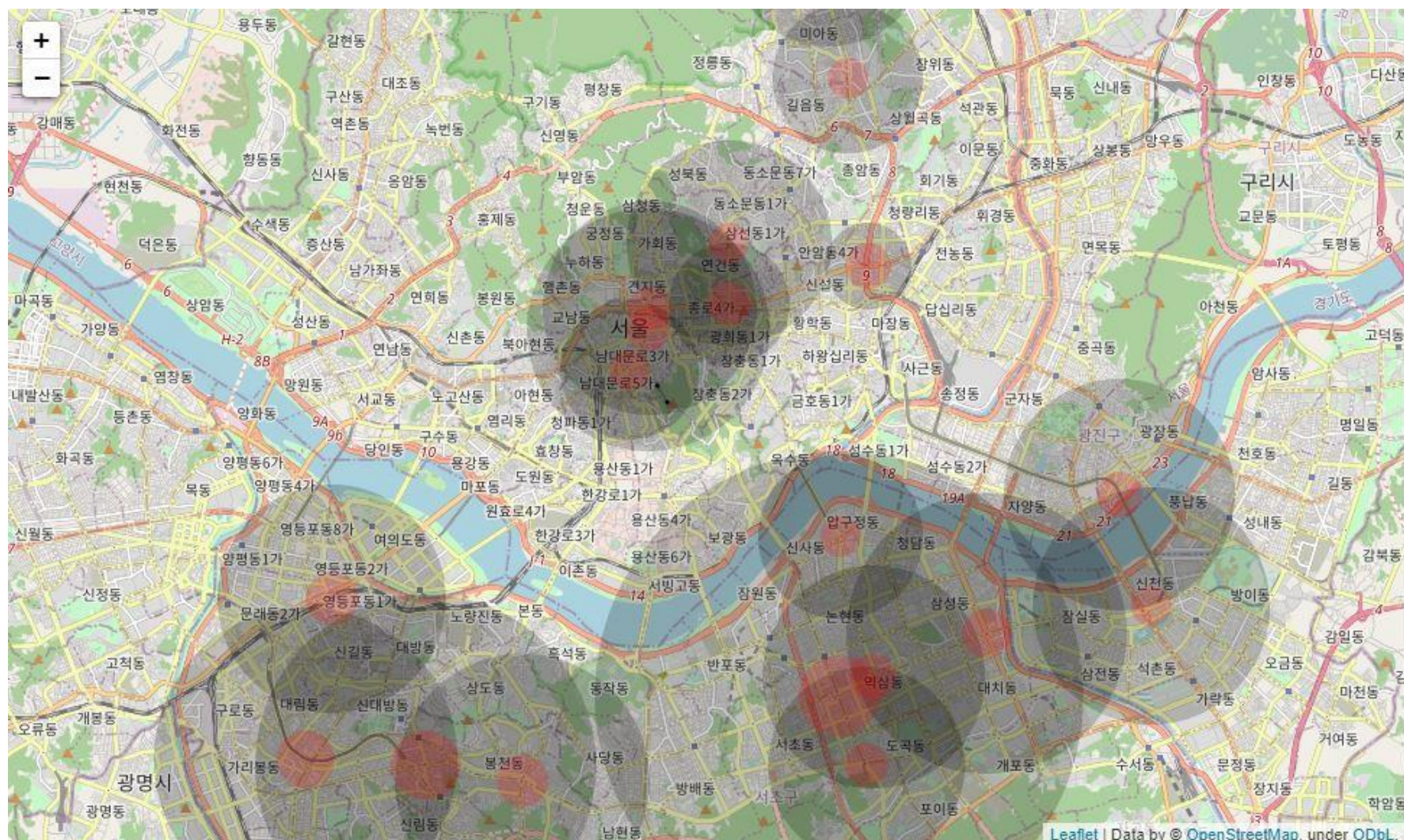


03

연도별 호선별 승하차인원 (heatmap)

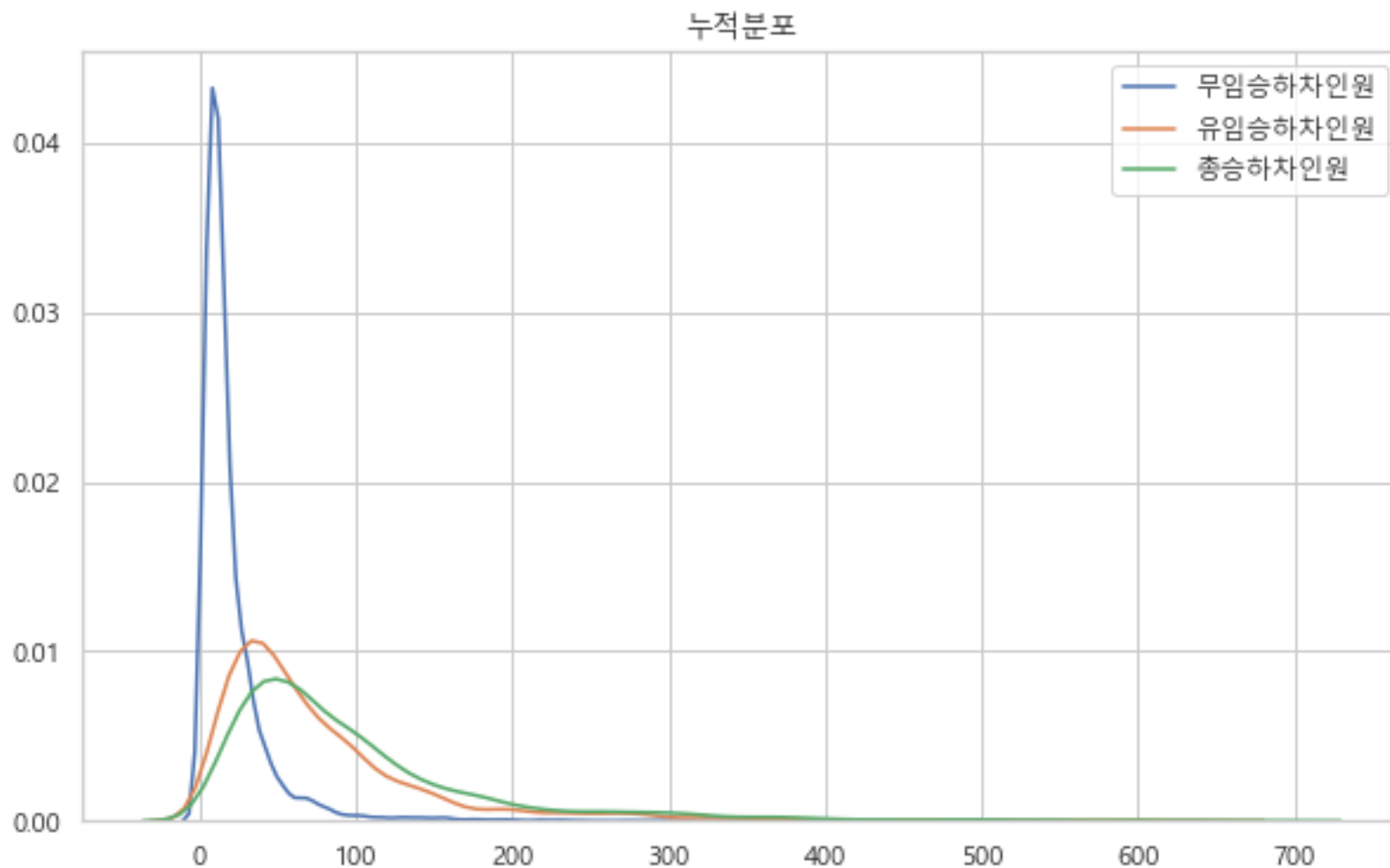


03

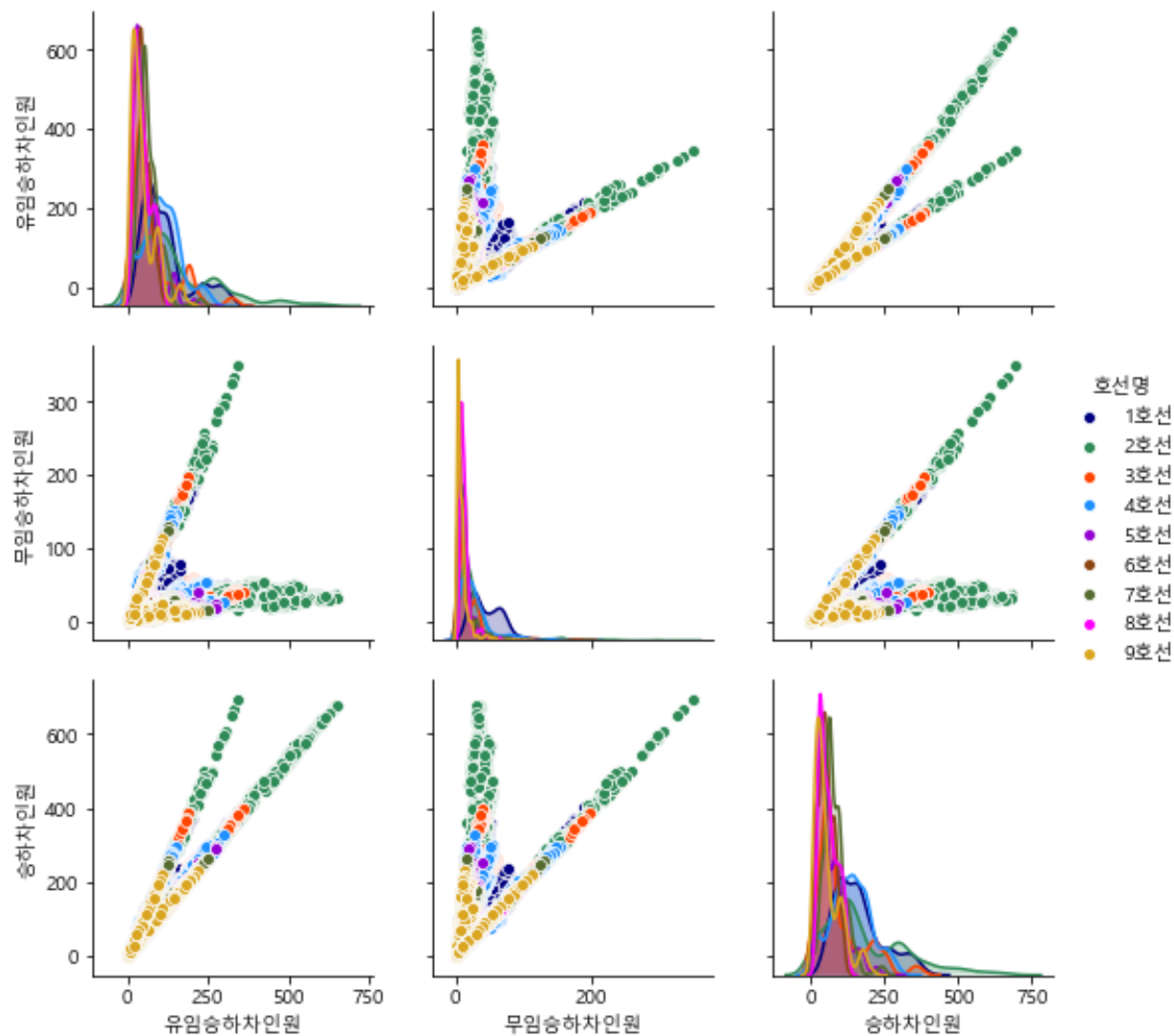


03

누적 분포 그래프

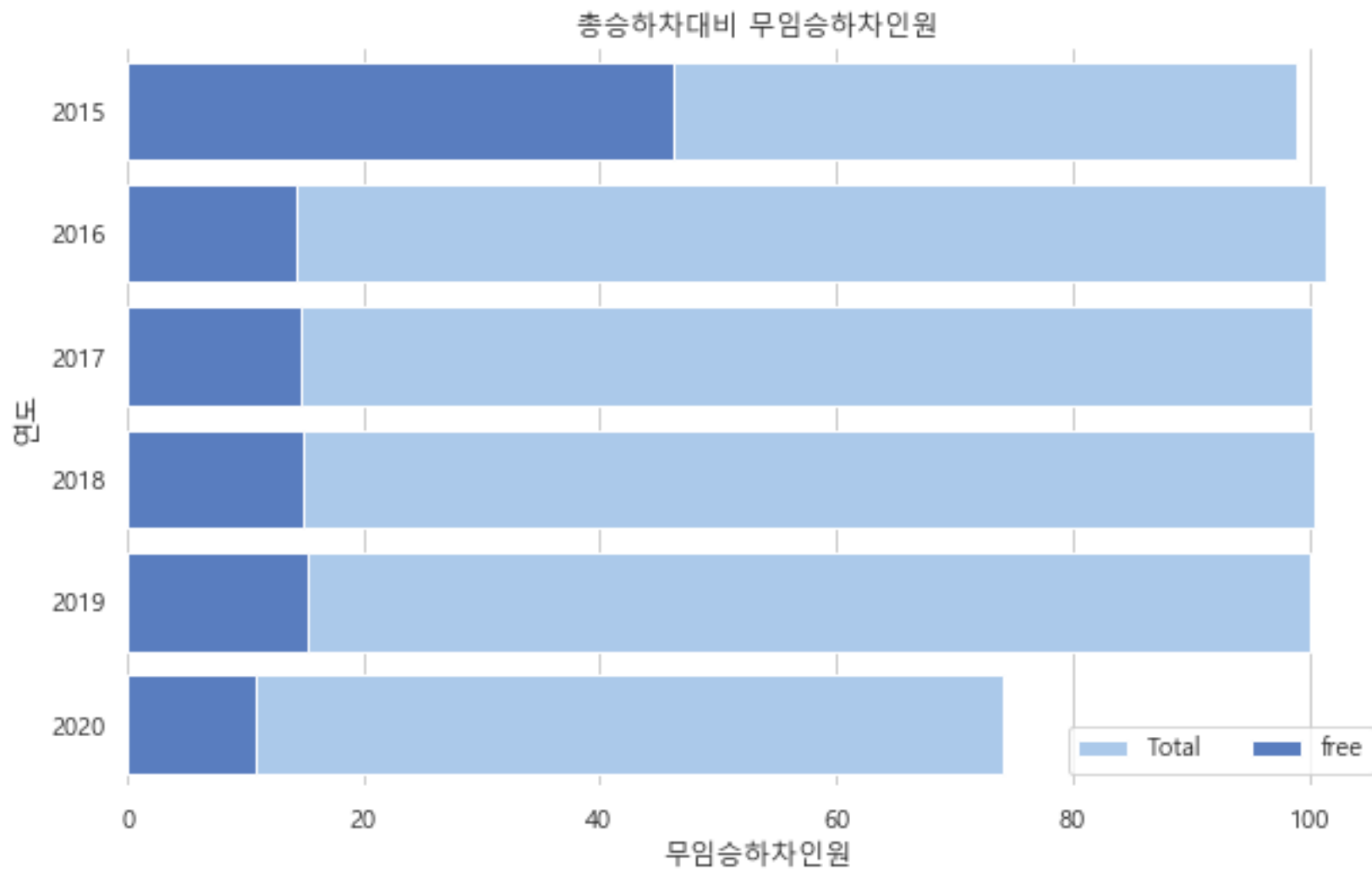


03



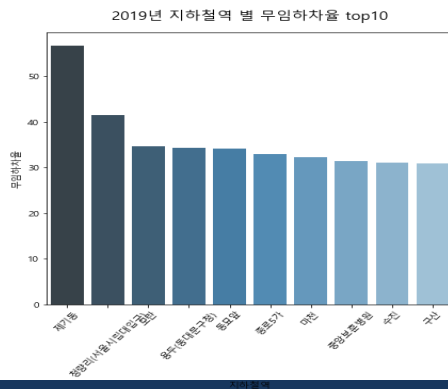
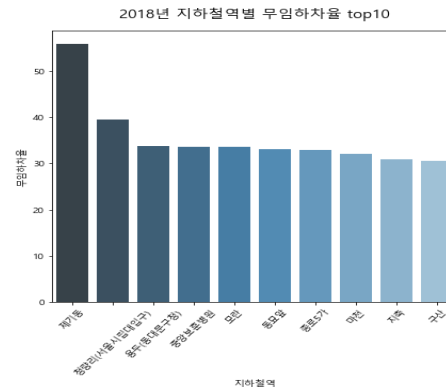
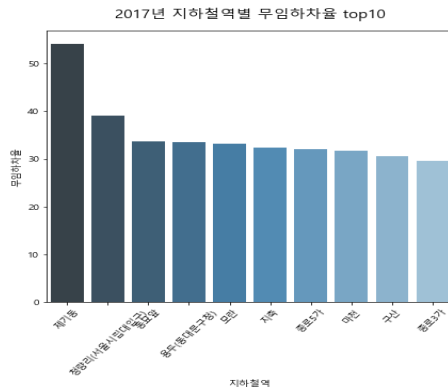
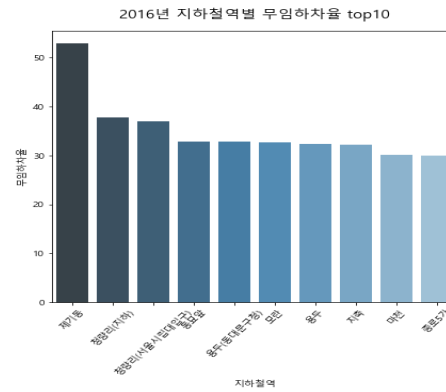
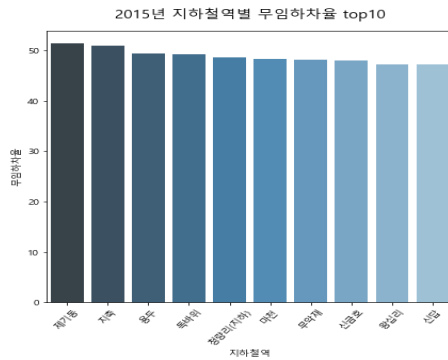
03

총 승하차 대비 무임승하차인원



03

연도별 무임하차율 상위역사 현황



최근 5개년 무임하차율 1위 역은?

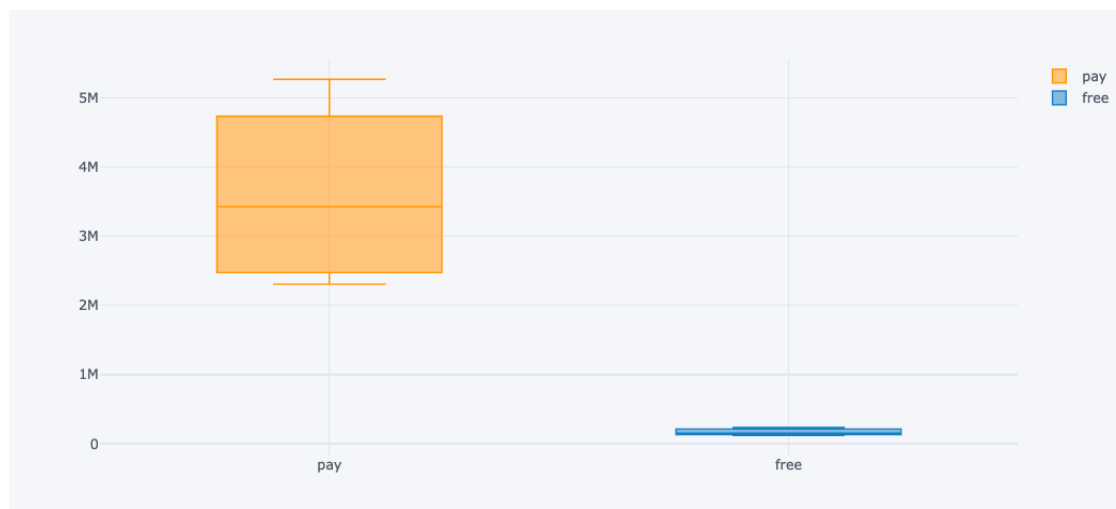
15년도 상위 10개의 역 모두
무임하차비율 50%에 이름

무임하차율 순위 상위역 모두 1호선

03

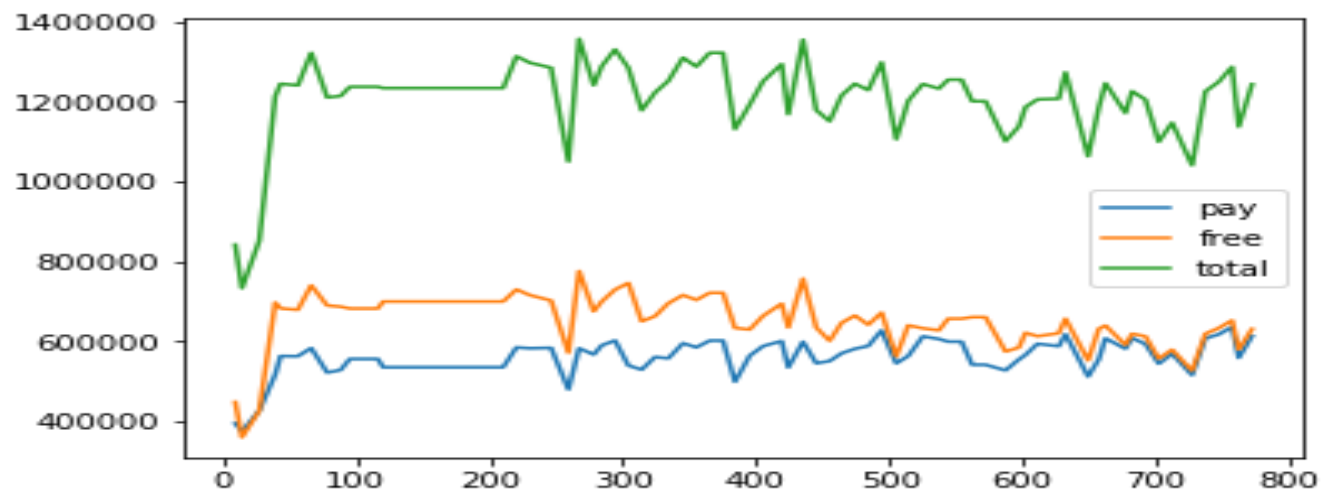


1호선 제기동

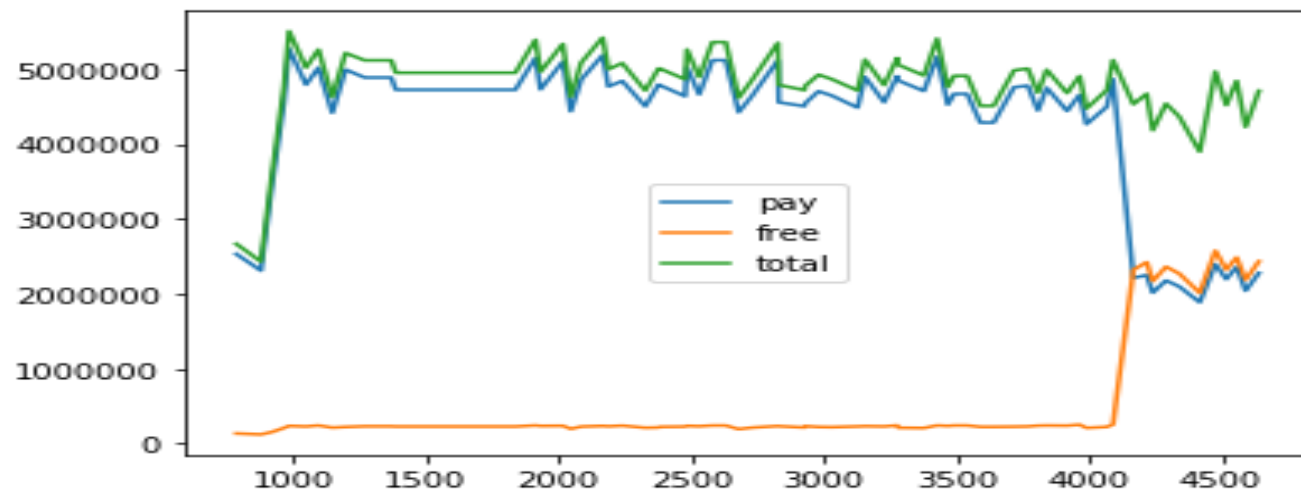


2호선 홍대입구

03



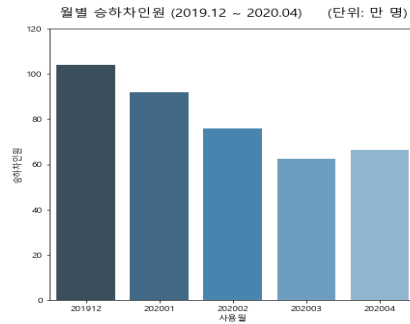
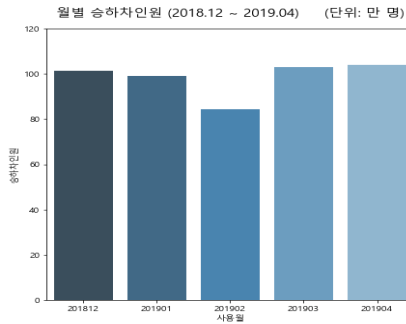
1호선 제기동



2호선 홍대입구

03

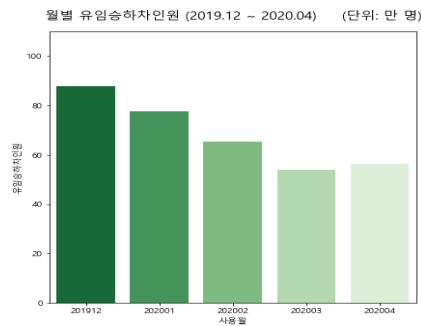
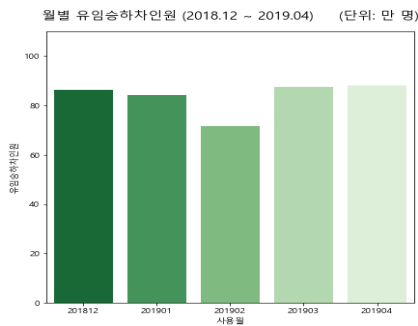
코로나 영향에 따른 이용객 수 변동 추이



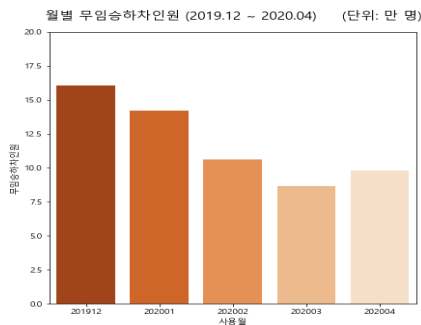
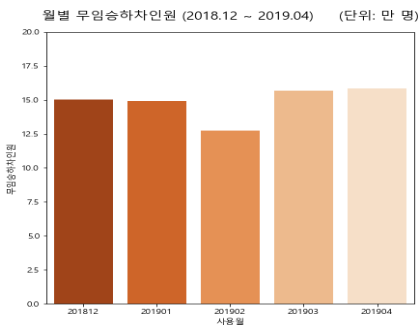
1. 승하차인원 변동 추이

좌) 2018.12 ~ 2019.04

우) 2019.12 ~ 2020.04



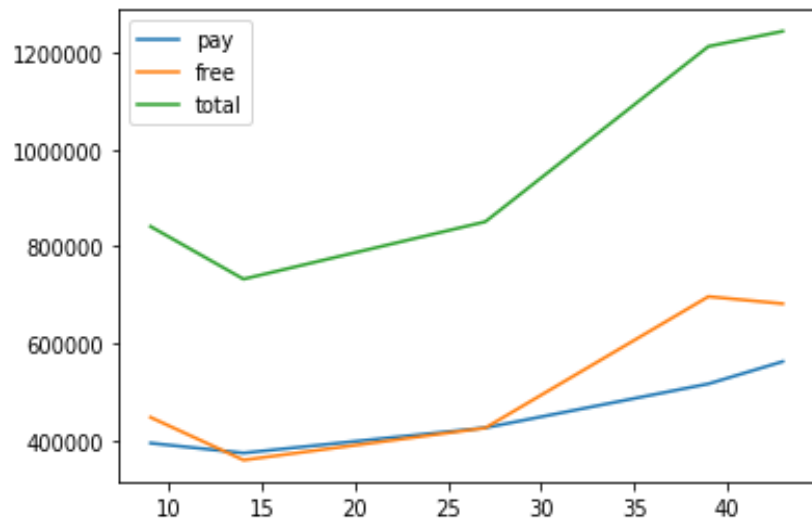
2. 유임승하차인원 변동 추이



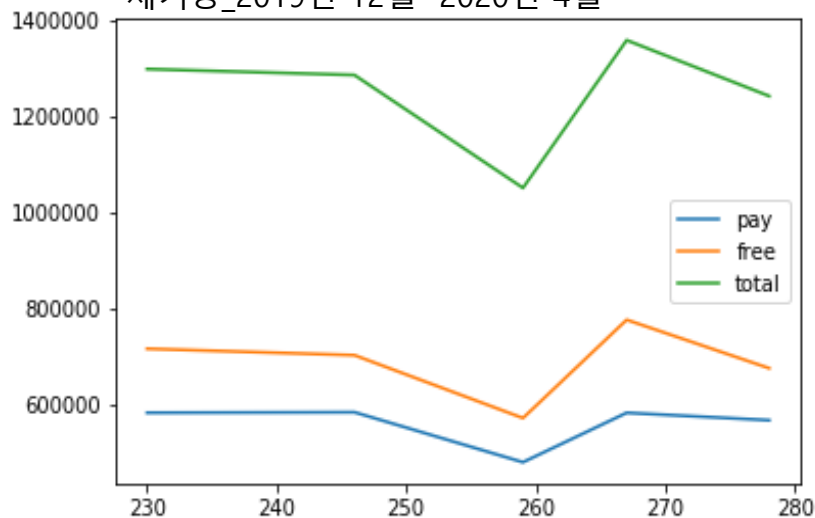
3. 무임승하차인원 변동 추이

03

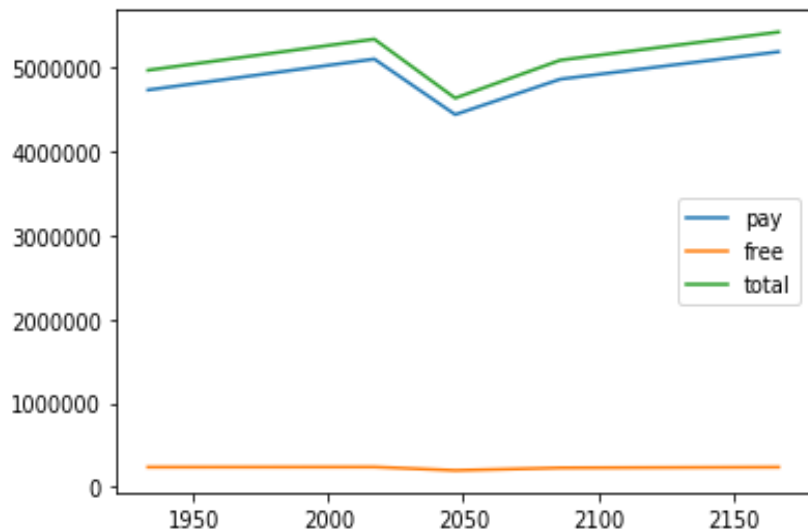
제기동_2018년1월 ~2019년 4월



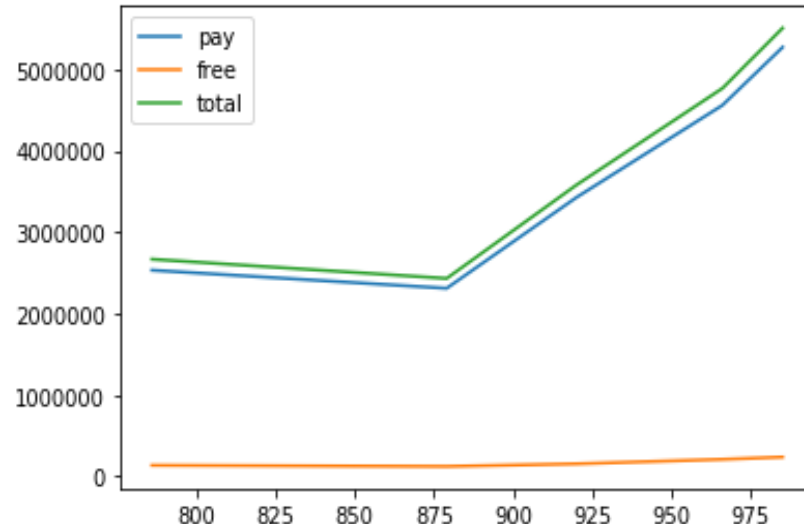
제기동_2019년 12월~2020년 4월



홍대입구_2018년1월 ~2019년 4월



홍대입구_2018년1월 ~2019년 4월



04

추세 확인

- 연도별 총 승하차인원 및 무임승하차인원의 변동
- 향후 이용승객의 변화 유추 가능

분포 확인

- 무임승차자 다이용 역사에 노인 편의시설 보강제안
- 주이용객 분류를 통해 역사내 입점 점포 선정에 반영 제안

코로나 영향

- 전년 동월대비 급격한 이용객 감소를 확인할 수 있음
- 특히 무임승객의 감소가 비교적 더 두드러짐

Reflection

- 시각화 프로젝트를 통해 데이터 가공, 시각화 자발 공부가능
- 간단한 데이터를 통해서도 분석에 한계를 느꼈다.

THANK
YOU

발 표 자 최 희 경