제주도 버스 승차인원 예측

홍지연, 임정민, 신보람

Contents

데이터 소개 및 분석 목적

데이터 소개

관측치 (obs.) : 415,423 개

Data (정류장 기준 탑승 정보)

date	Bus_ route_id	Station_ name	X6.7_ ride	X7.8_ ride	…	X6.7_ takeoff	…	X18.20_ ride	…
2019-09-01	4270000	제주 썬호텔	0	1	…	0	…	0	…

▷ 9월 한달동안 날짜별, 정류장별

- 오전 6시 ~ 오후 12시 승차인원
- 오전 6시 ~ 오후 12시 하차인원
- 오후 6시 ~ 오후 8시 승차인원

이 기록되어 있음.

▷ 결측치 존재 X

데이터 소개

관측치 (obs.) : 1,548,759 개

Bus_bts (이용자 기준 탑승 정보)

User_ card_id	Bus_ route_id	Geton_ date	Geton_ time	Geton_ station_ name	Getoff_ date	Getoff- time	Getoff_ Station_ name	User_ category	User_ count
4.33028 9e+15	4270000	2019- 09-01	07:48:24	제주 썬호텔	NA	NA	NA	1	1

▷ 버스카드별 승하차 정보가 기록되어 있음. (단, 탑승시간대가 오전 6시~낮 12시인 경우만 존재)

▷ 하차 태그를 안 하는 경우 결측치가 존재

▷ user_category (승객 구분) : 01-일반, 02-어린이, 04-청소년, 06-경로, 27-장애 일반, 28-장애 동반, 29-유공 일반, 30-유공 동반

▷ user_count : 해당 버스카드로 계산한 인원 수

분석 목적

- 2017년 한국은행 제주본부에 따르면 제주도 일부 지역은 교통체증이 서울보다 심각.
- 퇴근시간대 승차인원을 예측함으로써 제주도 버스의 효율적인 운행이 필요.

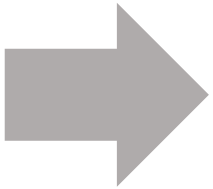
출근 시간대 승차인원

승객 유형

버스 노선 특징

정류장 특징

…



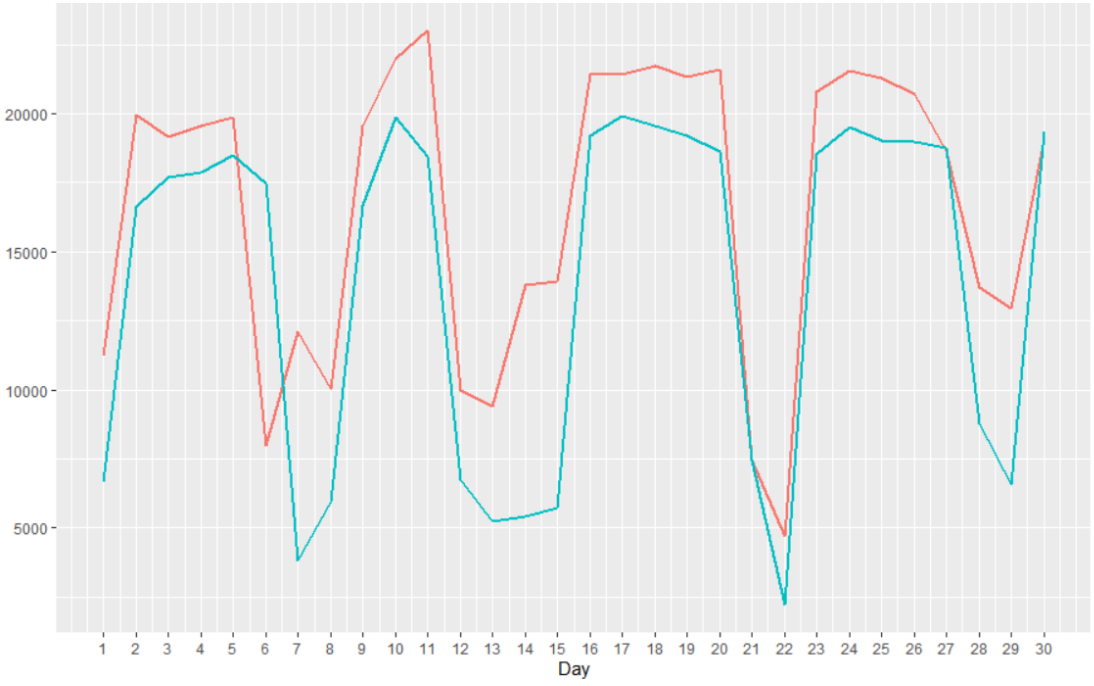
퇴근 시간대 버스 승차 인원 예측

ex. 9월 1일, 출근 시간대 승차정보가 주어졌을 때
퇴근 시간대 '제주썬호텔' 정류장에서
'4270000' 노선 버스에 승객이 몇 명 탈까?

데이터 전처리

변수 정리 출근시간 승차인원과 퇴근시간 승차인원의 연관성

6~8시 승차인원과 18시~20시 승차인원

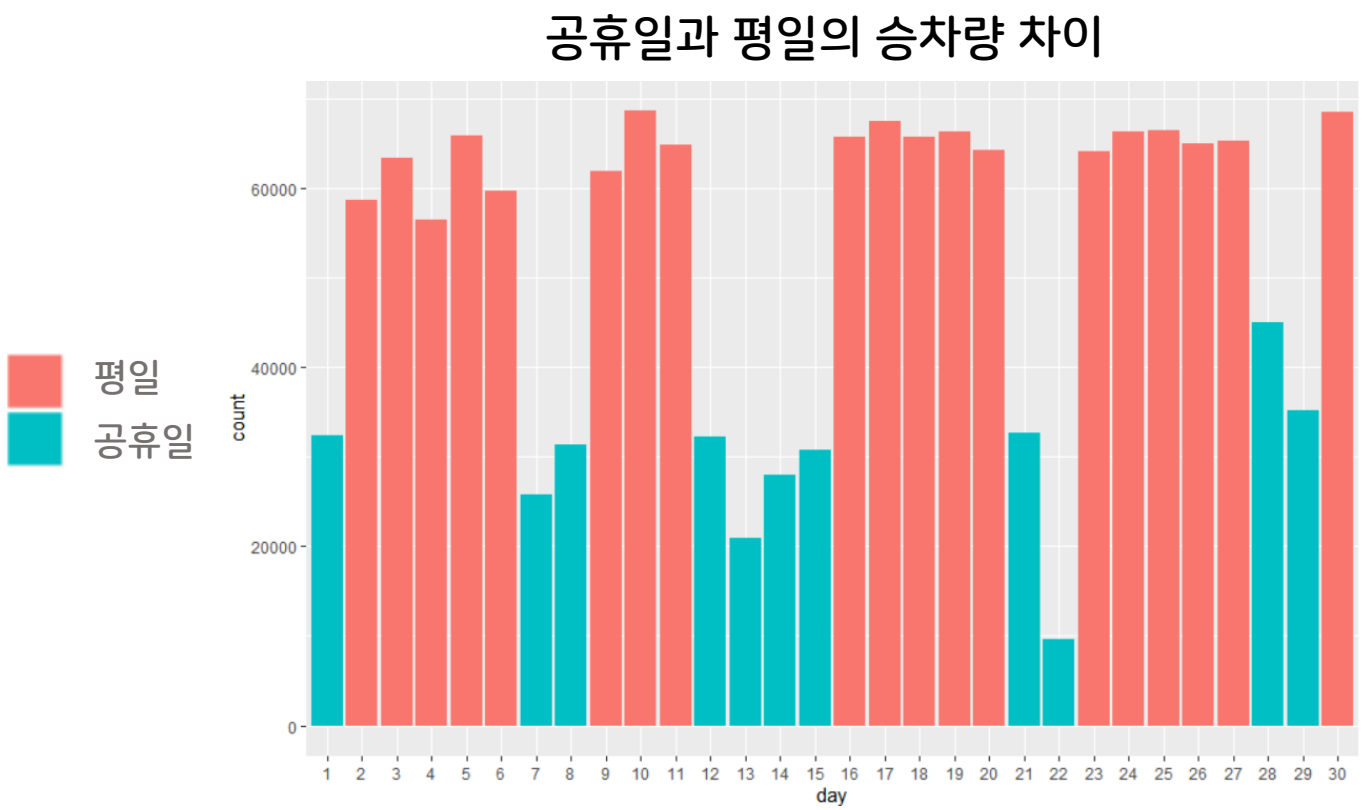


6~8시 승차인원과 18시~20시 승차인원의
평일 패턴이 유사함 -> 평일 출근시간 버스 이용고객은
대체적으로 퇴근시간에도 버스를 이용함

▷ 한 시간 단위로 기록되어 있는 승하차 인원을
3시간 단위로 묶음.

ex. 6~7시 승차인원 + 7~8시 승차인원 + 8~9시 승차인원
➡ 6~9시 승차인원 (Geton_6.9)

변수 정리 출근시간 승차인원과 퇴근시간 승차인원의 연관성

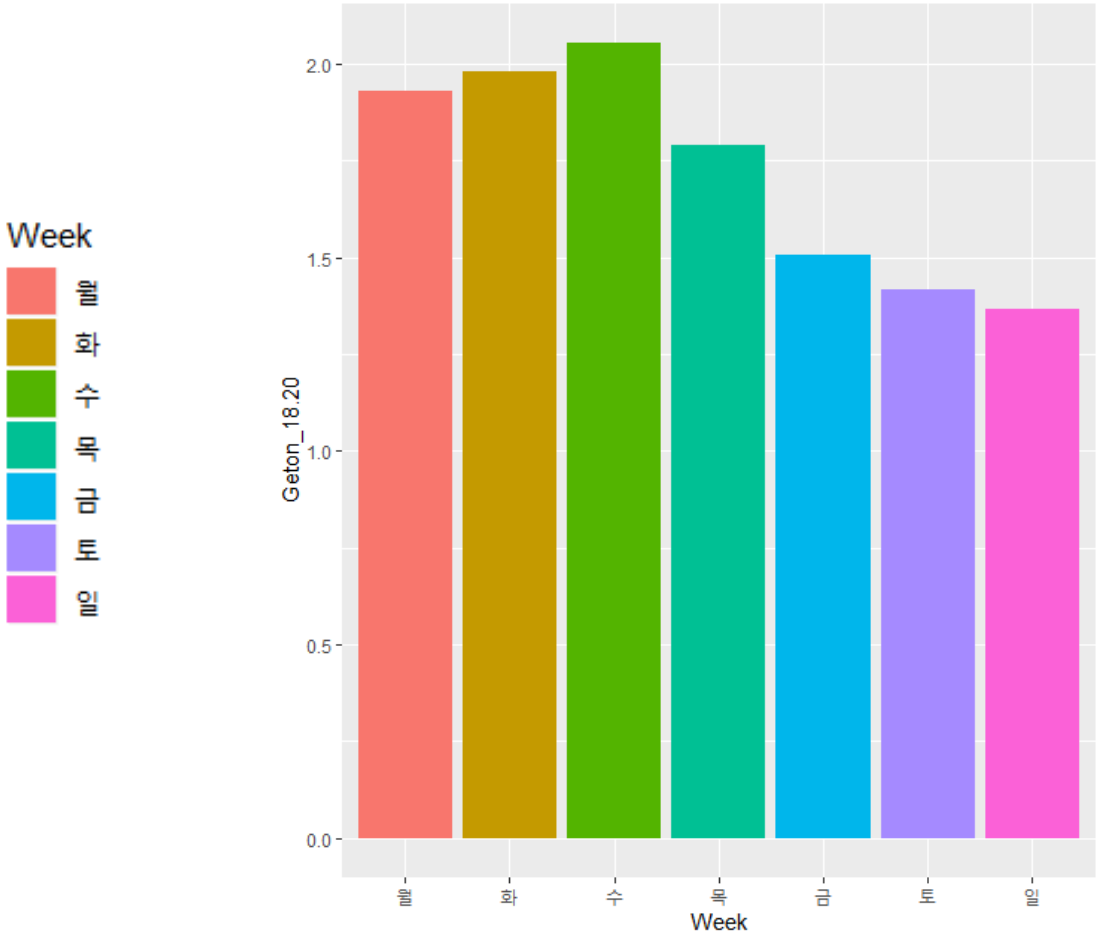


평일 > 공휴일 (하루 총 승차량)

변수 생성 : Holiday

(공휴일이면 1, 평일이면 0)

변수 정리 요일별 퇴근시간 승차인원



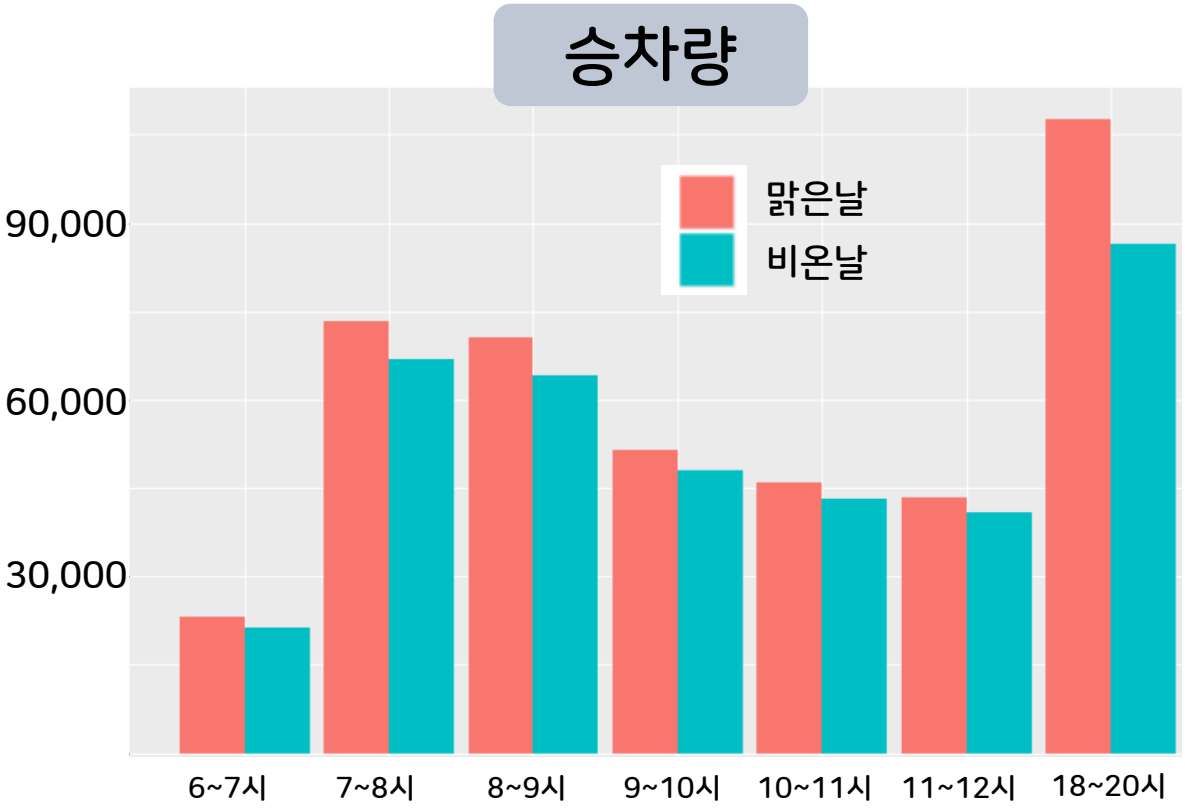
요일별로 승차 인원에 차이가 있음.

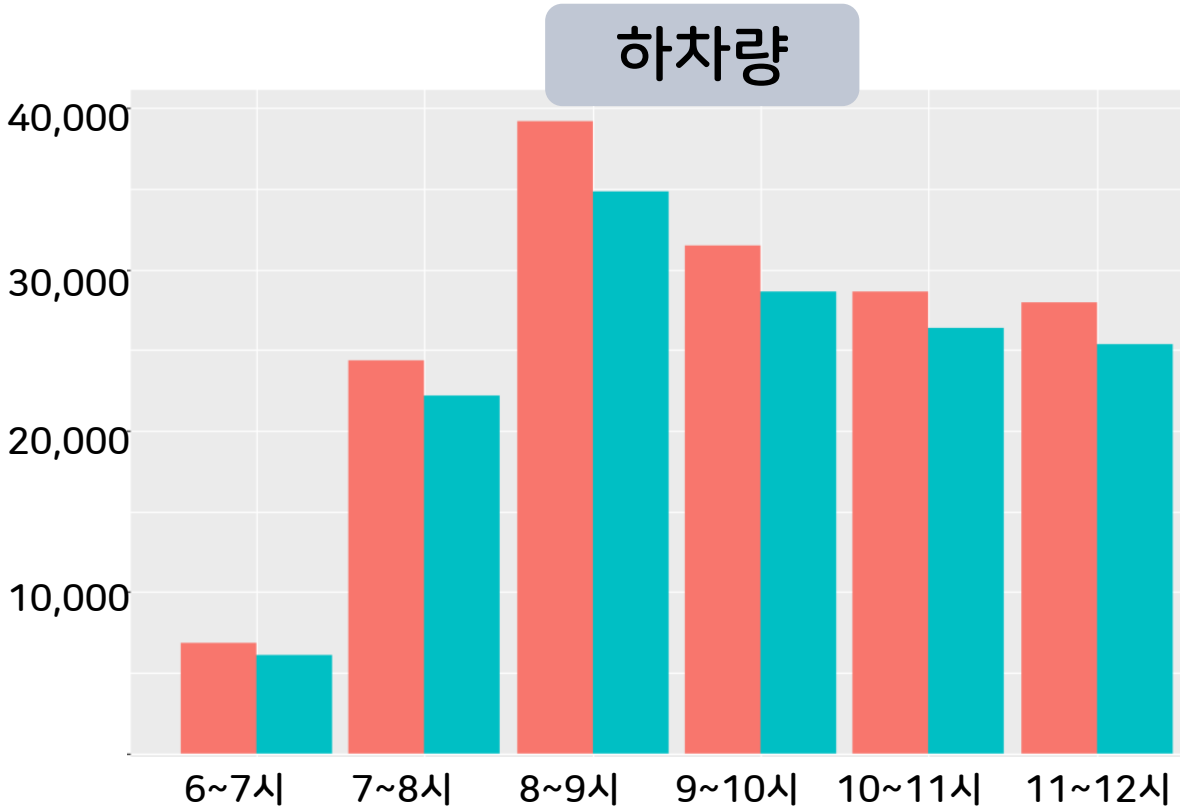
변수 생성 : Mon , Tue , Wed ,

Thu , Fri , Sat , Sun

변수 정리 맑은 날과 비 온 날의 총 이용량 차이

※ 동일 노선을 비교함





매일 비가 온 9월 첫째주와 매일 맑았던 9월 셋째주의 시간대별 승차량 비교 (평일)

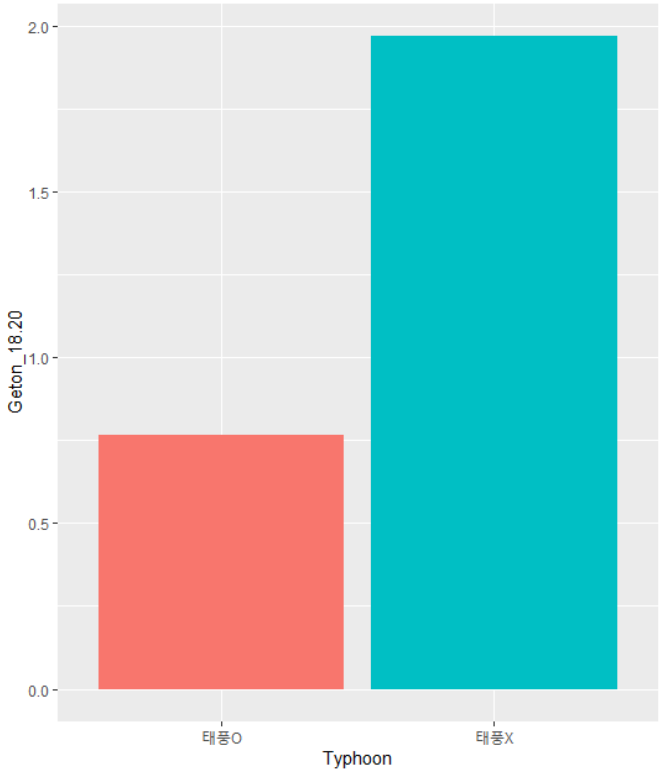
9월 한달간 집중호우와 태풍의 영향으로 비가 오는 날, 전지역의 강수량 차이가 크지 않았음.

변수 생성 : prcp (비가 왔으면 1, 안 왔으면 0)

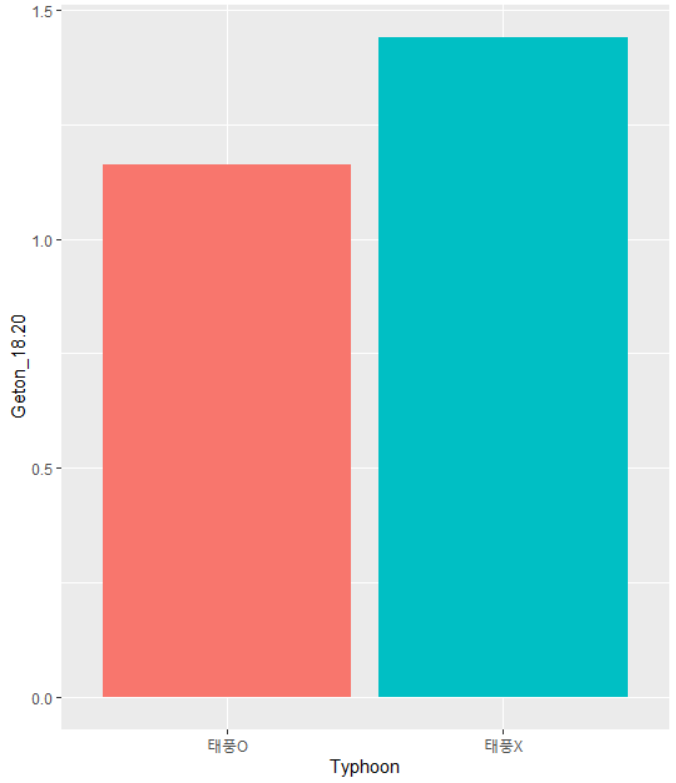
※ 9월 6, 7, 21, 22일 제주도에 태풍이 옴.

변수 정리 태풍이 왔을 때와 안 왔을 때의 승차량 차이

평일



공휴일

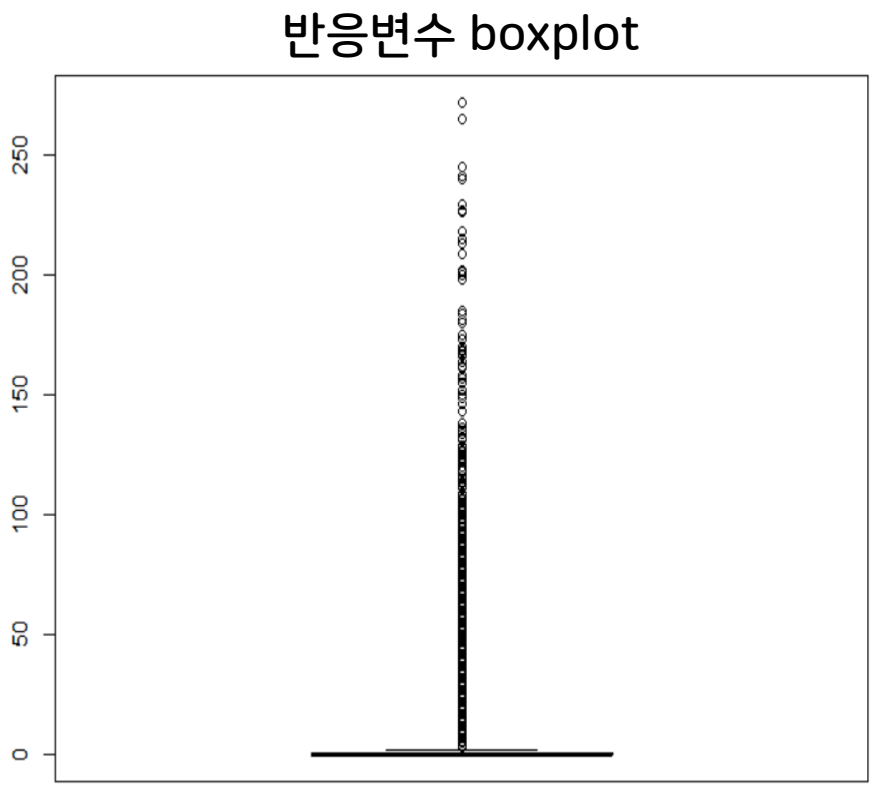


Typhoon
태풍O
태풍X

태풍이 안 왔을 때 > 태풍이 왔을 때

변수 생성 : Typhoon (태풍이 왔으면 1, 안 왔으면 0)

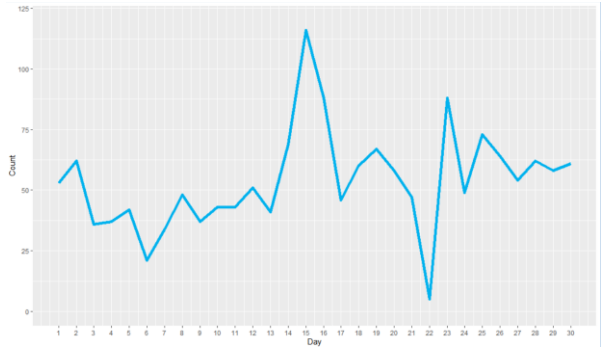
변수 정리 이상치가 많은 승차량 변수



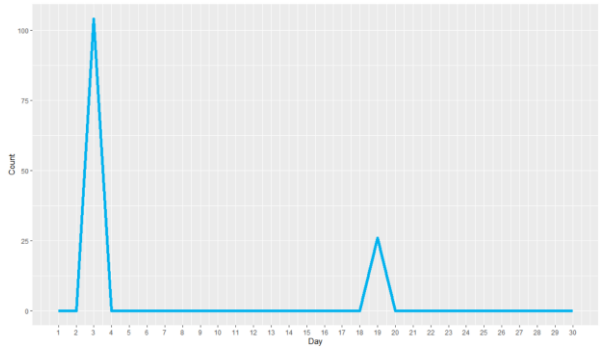
- 반응변수인 퇴근시간의 승차량에 **이상치가 많음.**
- 반응변수의 64%가 0명에 해당함.
- 심각히 치우친 분포이며 불균형 데이터임.

변수 정리 이상치가 많은 승차량 변수

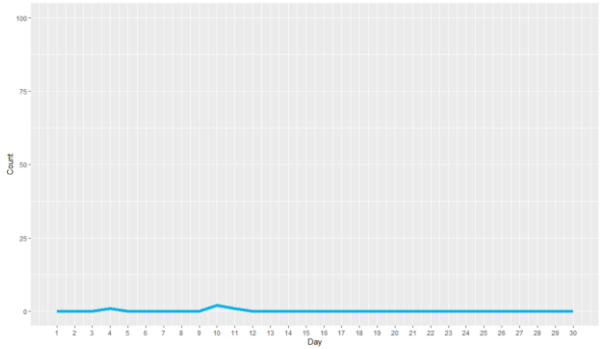
y축 :
퇴근시간
승차인원



유형1



유형2



유형3

x축 : 날짜

- **(노선+ 정류소) 그룹별** 한달 간 승차인원의 특징을 파악함
  - ✓ 유형 1)
    - 대체적으로 승차인원 많음
    - 요일 별 패턴 보임
  - ✓ 유형 2)
    - 거의 모든 날 승차인원 0명
    - 가끔 승차인원 많은 날 존재
  - ✓ 유형 3)
    - 대체적으로 승차인원이 적음

변수 정리 이상치가 많은 승차량 변수

이상치 대체? 삭제?

- 범주화하면?
  - (노선+정류장) 그룹마다 승차인원 특성이 달라 기준이 모호
- (노선+정류장) 그룹마다 3Q + 1.5*IQR 값으로 이상치 대체?
  - (유형2) 같은 패턴은 예)108명 -> 0명으로 대체됨.
- 0명이 기록이 잘못된 것 아닐까?
  - 데이콘 확인 결과, 기록에 이상 없음
  - 현금 손님만 있었다면 승차인원이 기록되지 않았을 수도.

대신 승차량에 따라

정류장 군집화

데이터 전처리

변수 정리 승차량에 따른 정류장 군집화

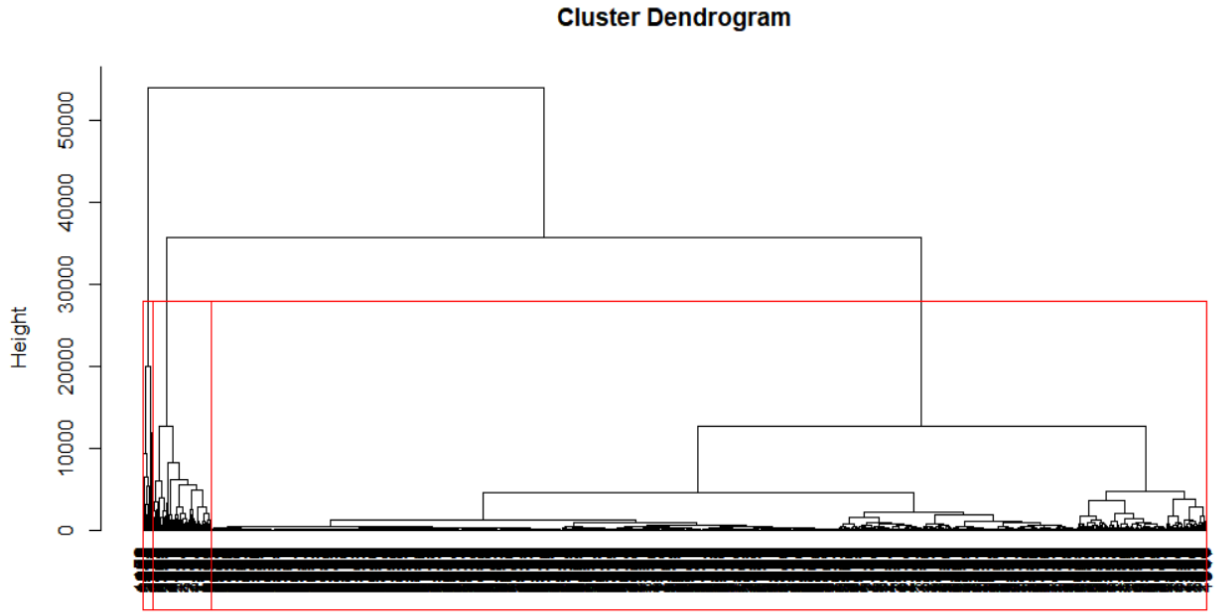
2570개의 정류소를 **시간대별 총 승차량**에 따라 군집화

Ward 결합방식을 사용한

계층적 군집분석(hierarchical clustering)결과,

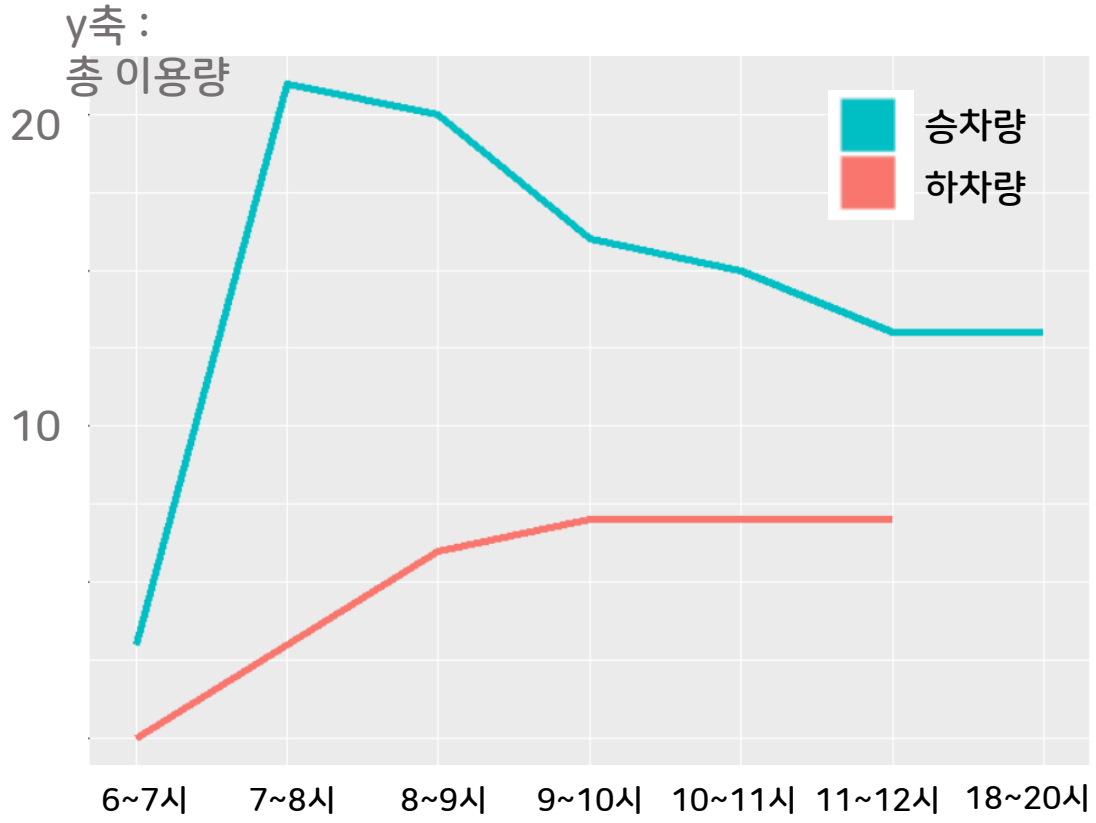
총 3개의 그룹으로 군집화 되는 과정을 확인하였다.

- ✓ 그룹1) 대체적으로 승차량이 적은 정류소 (2407개)
- ✓ 그룹2) 출근시간에 승차량이 집중된 정류소 (140개)
- ✓ 그룹3) 대체적으로 승차량이 많은 정류소 (23개)

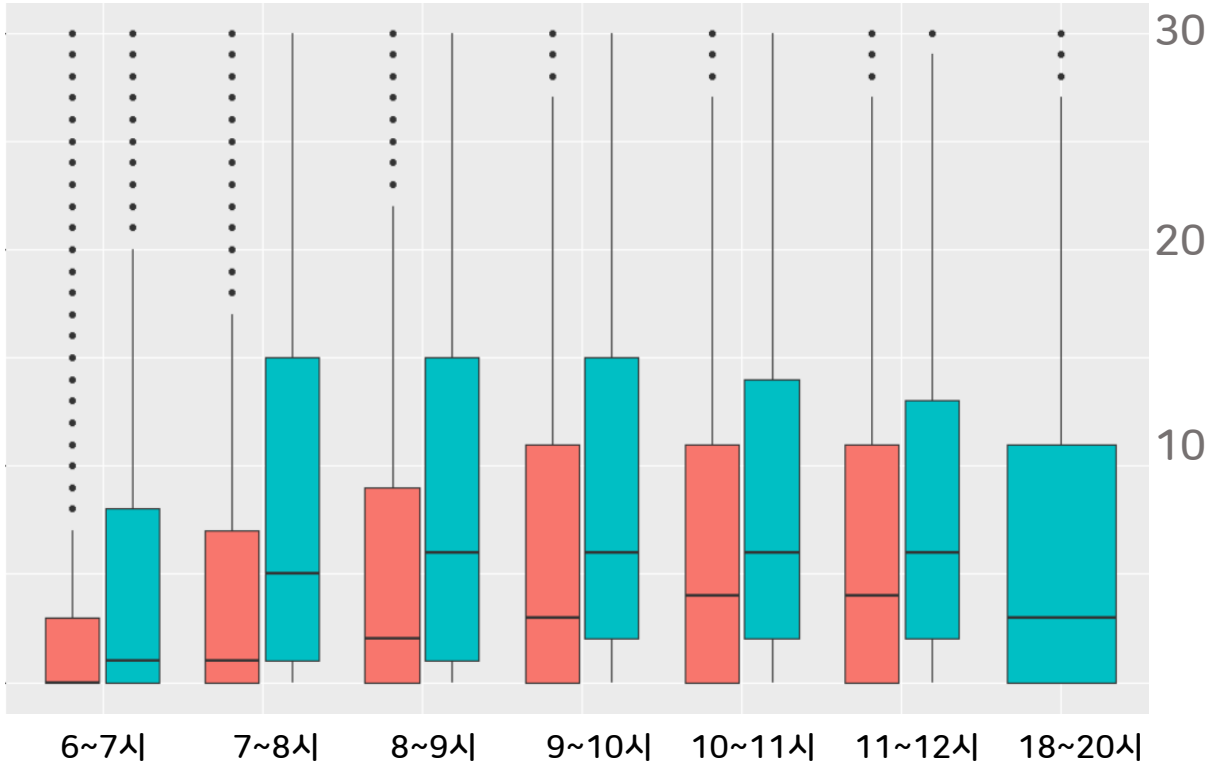


※ Average Silhouette width 실루엣 = 0.85
(4개 그룹일때 0.85, 5개 그룹일때 0.67)

변수 정리 승차량에 따른 정류장 군집화



< 그룹1 정류장 시간대별 이용량 중간값 >

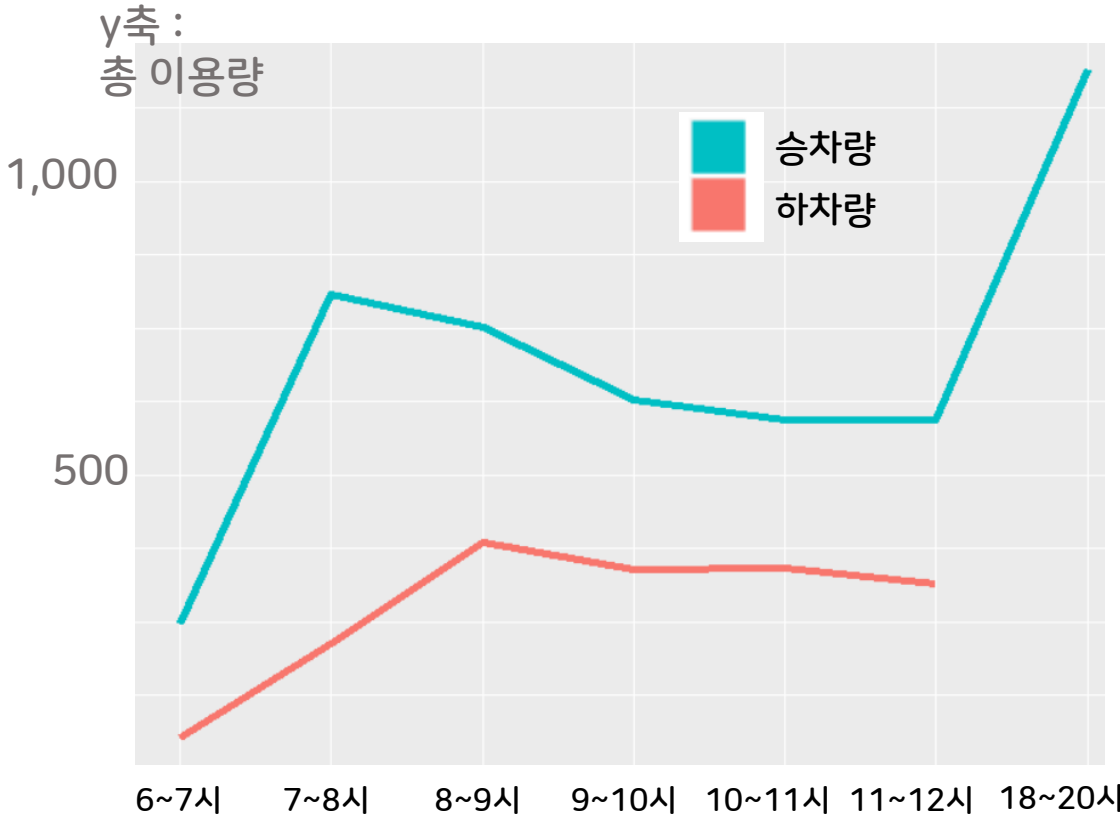


< 그룹1 정류장 시간대별 이용량 분포 >

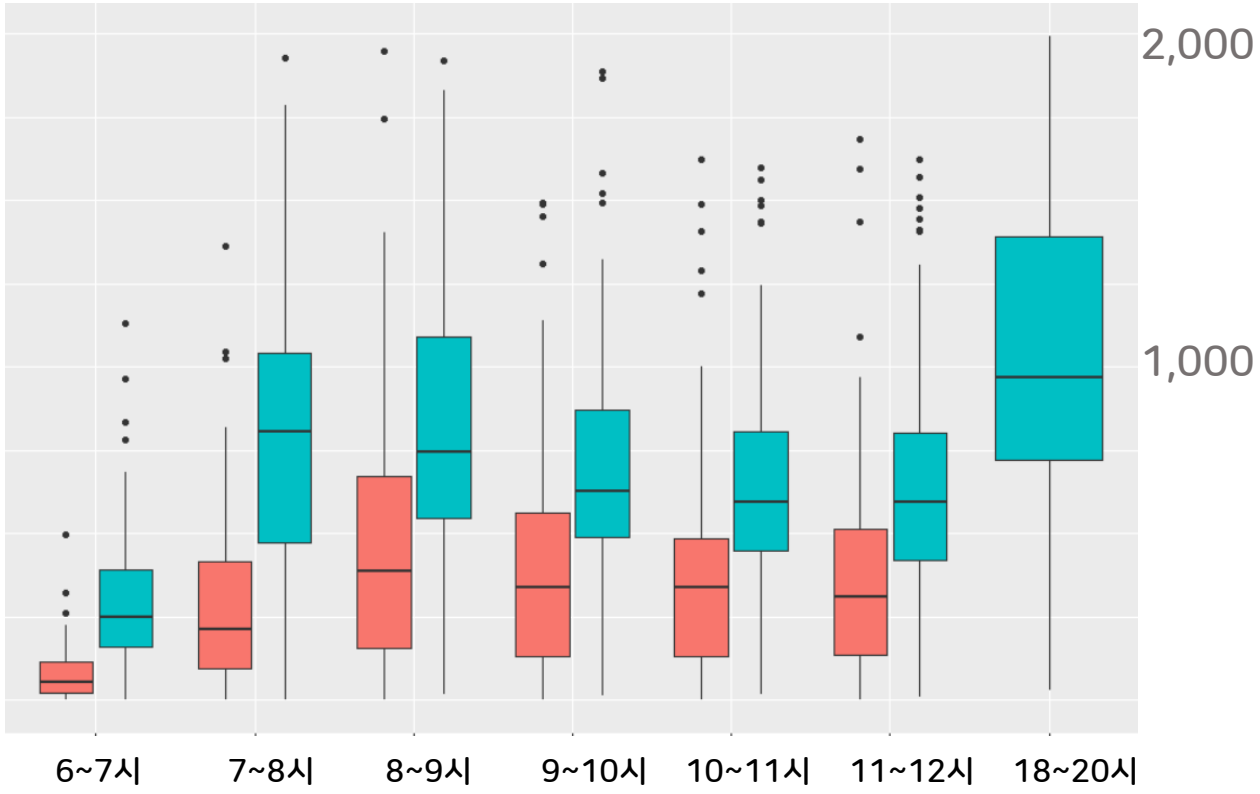
그룹1 : 시간대별 승차량이 대체적으로 적음

라벨링 : station_g2 = 0, station_g3 =0

변수 정리 승차량에 따른 정류장 군집화



< 그룹2 정류장 시간대별 이용량 중간값 >

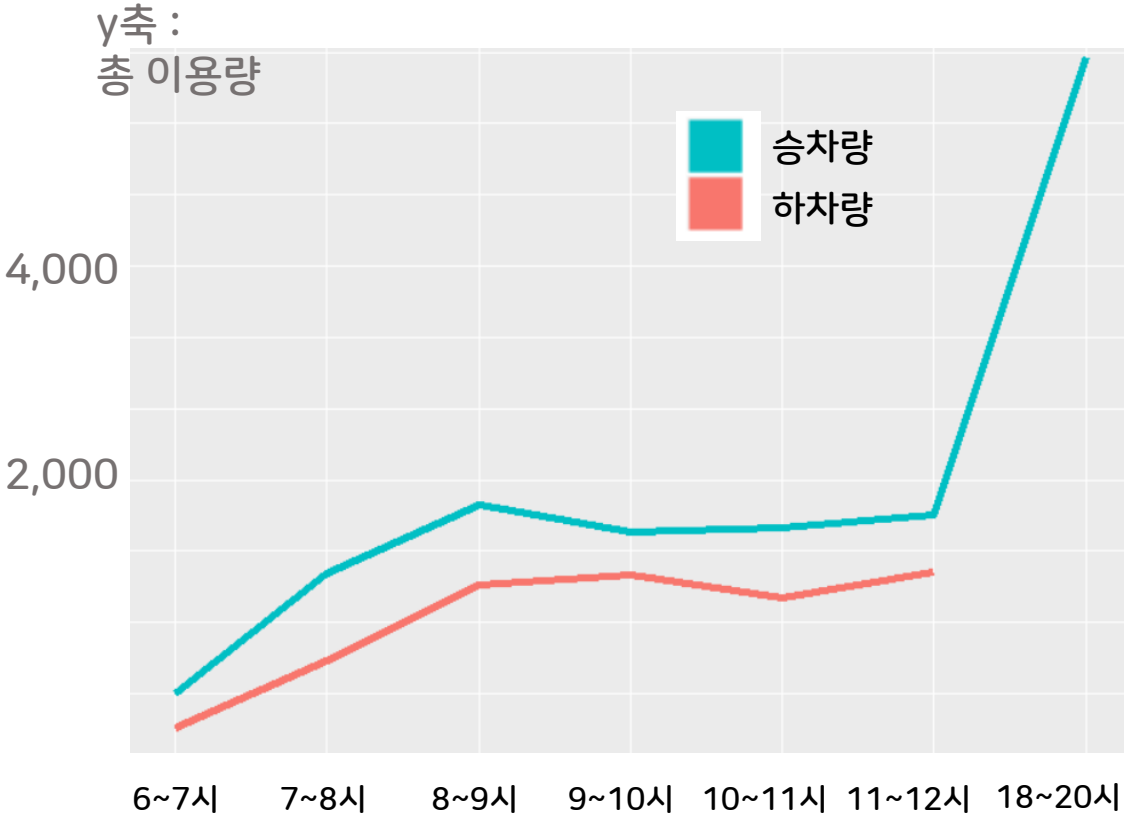


< 그룹2 정류장 시간대별 이용량 분포 >

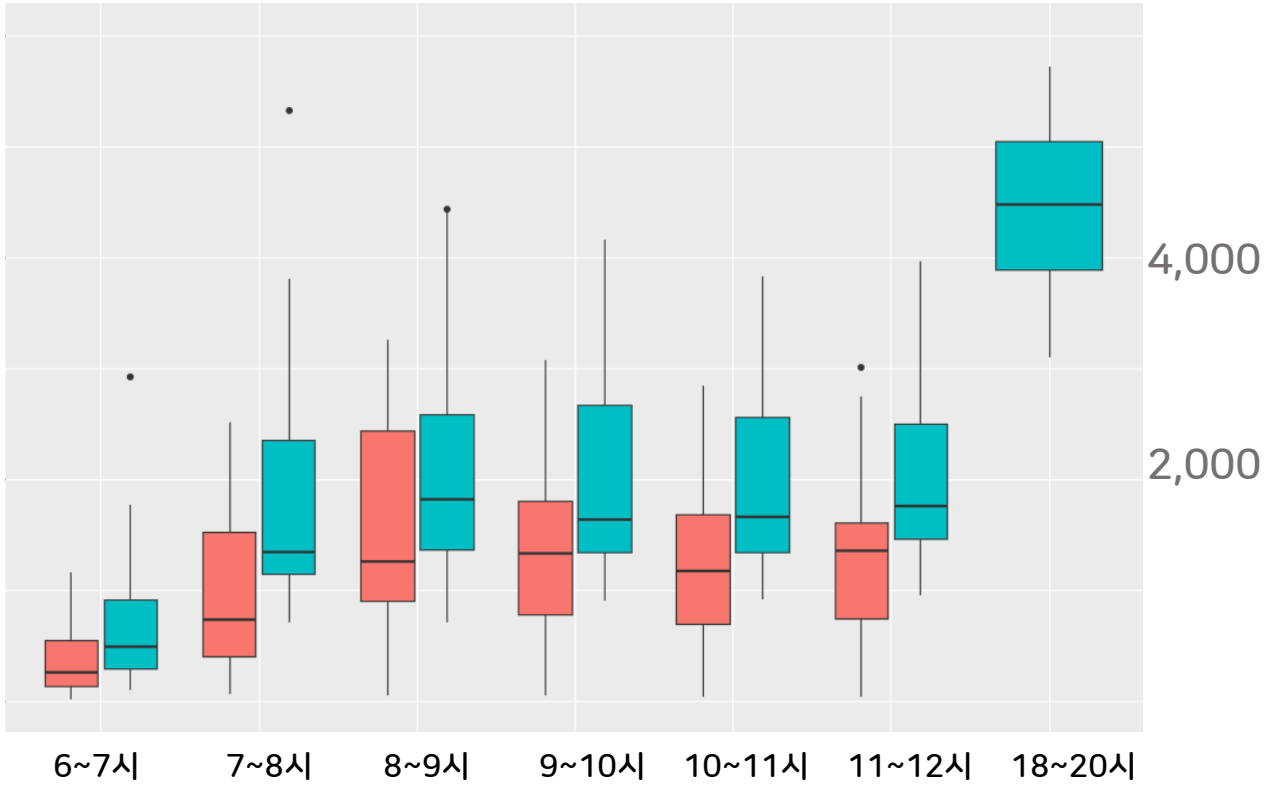
그룹2 : 출근시간대에 승차량이 집중됨

라벨링 : station_g2 = 1, station_g3 = 0

변수 정리 승차량에 따른 정류장 군집화



< 그룹3 정류장 시간대별 이용량 중간값 >



< 그룹3 정류장 시간대별 이용량 분포 >

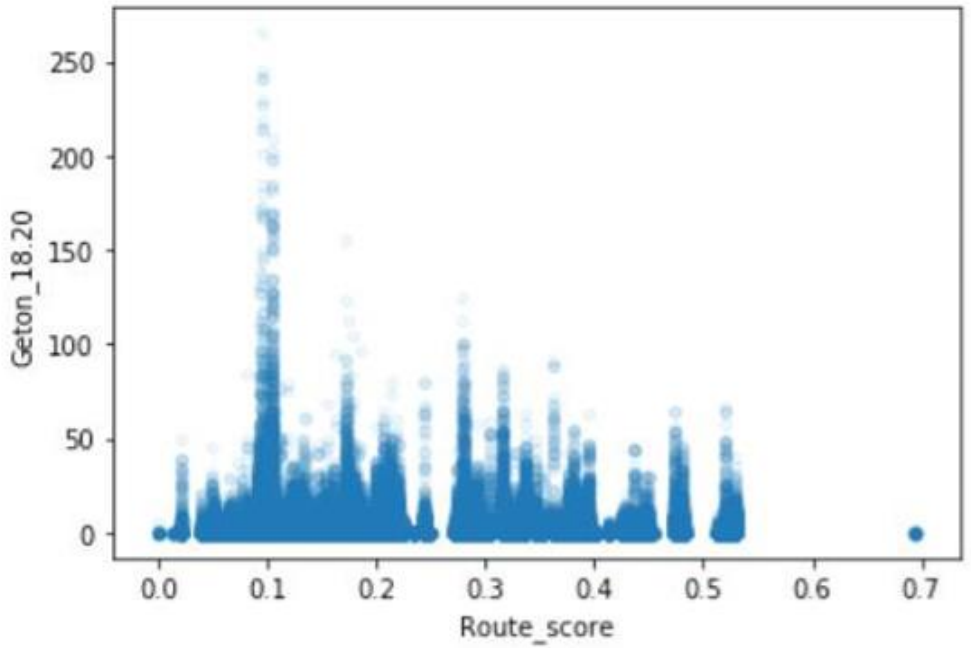
그룹3 : 시간대별 승차량이 대체적으로 많음

라벨링 : station_g2 = 0, station_g3 = 1

변수 정리 노선의 특성을 나타내는 변수

▷ 특정 노선에 대해 정류장이 흩어져 있는 정도를 나타내는 route_score 생성.

$$\text{route\_score} = \sqrt{(\text{위도 차이 max})^2 + (\text{경도 차이 max})^2}$$



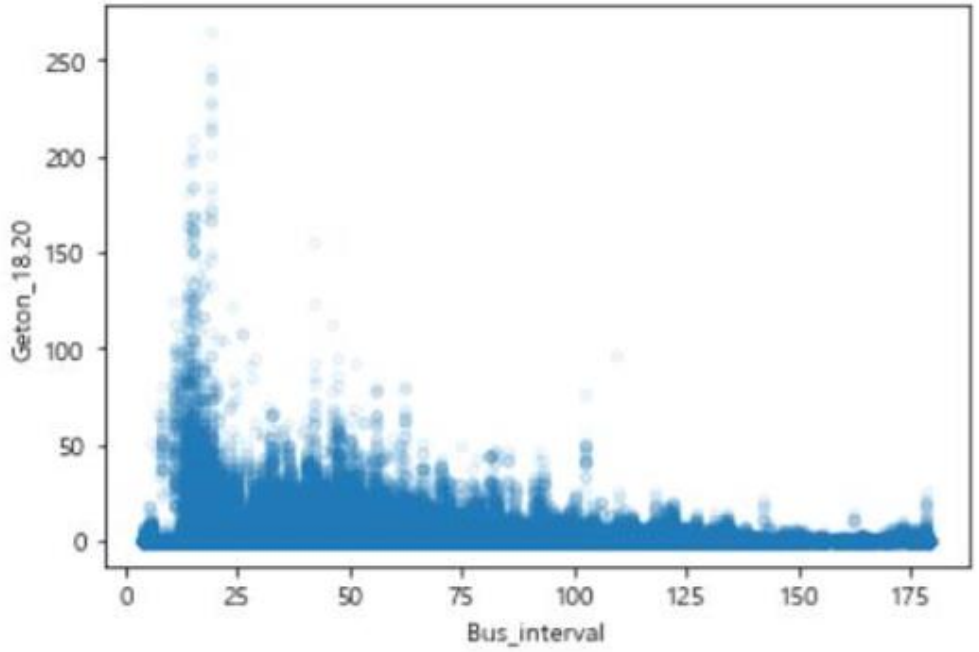
노선 점수가 0.1일 때
(짧은 노선일 때)
승차인원이 많음.

변수 정리 노선의 특성을 나타내는 변수

▷ 버스 배차 간격을 계산한 bus_interval 생성.

하루 동안 노선+정류장별로 배차간격의 평균을 구해 정류장마다 기록함.

승차 기록이 한 개인 경우 배차 간격을 구하지 못함. → 행 삭제

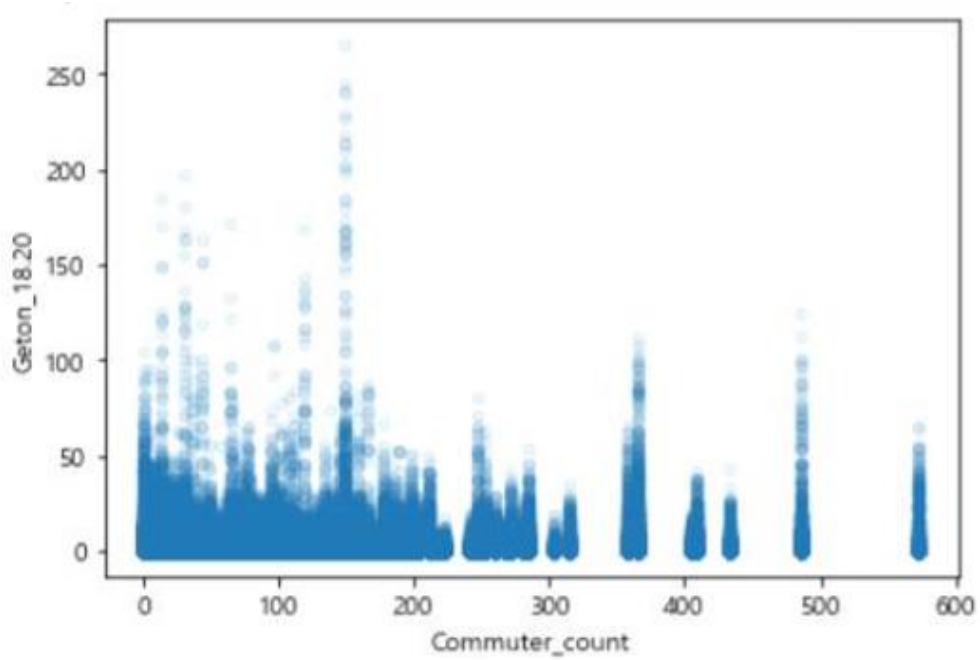


배차 간격이 10~20분일 때
승차인원이 가장 많음.

변수 정리 정류소의 특성을 나타내는 변수

▷ 통근자 수를 나타내는 commuter_count 생성.

9월 평일에 동일한 노선+정류장을 10번 이상 이용한 이용자를 제주도민 통근자라고 판단함.



통근자 수가
약 150명 존재하는 정류장에서
승차인원이 가장 많음.

변수 정리 승하차 승객의 유형을 나타내는 변수

▷ 승객 유형별 오전시간 승하차 인원을 계산한
Type_general_geton (getoff), Type_others_geton (getoff) 생성.

Data (정류소 기준, 6~12시)와 bus_bts (이용자 기준, 6~12시)를 merge하면서 하차 인원에서 결측치 발생함.

1) 하차 태그를 안 한 경우
2) 오전 12시 이후에 하차한 경우 : data에는 승차 기록만 남고, bus_bts에는 승하차 기록 모두 남음
3) 오전 6시 이전에 승차한 경우 : data에는 하차 기록만 남고, bus_bts에는 기록 안됨.

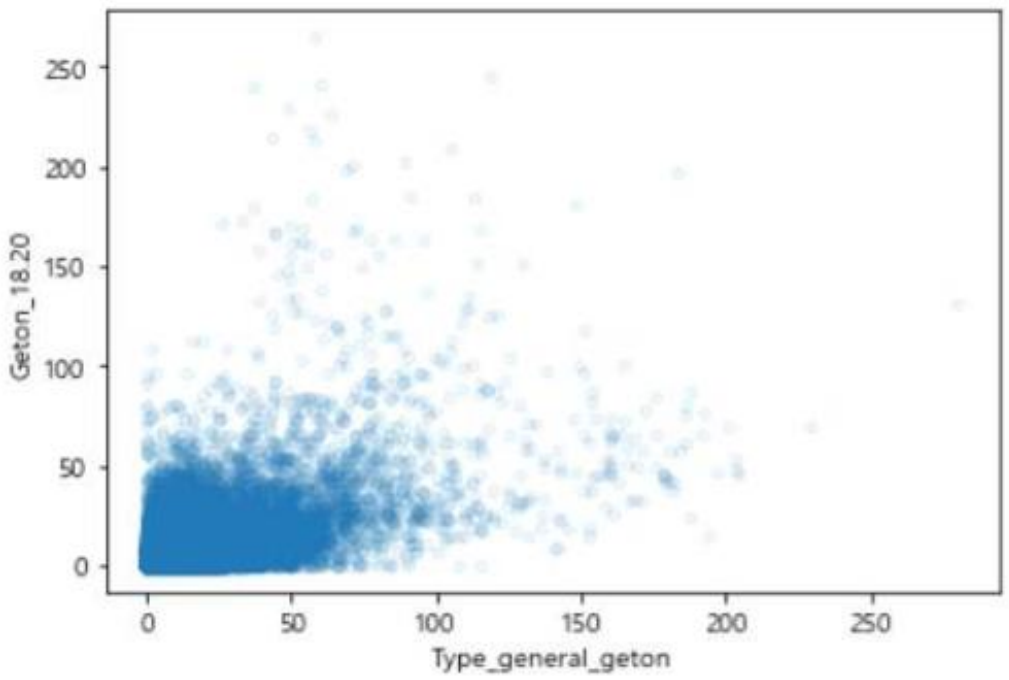
결측치 처리

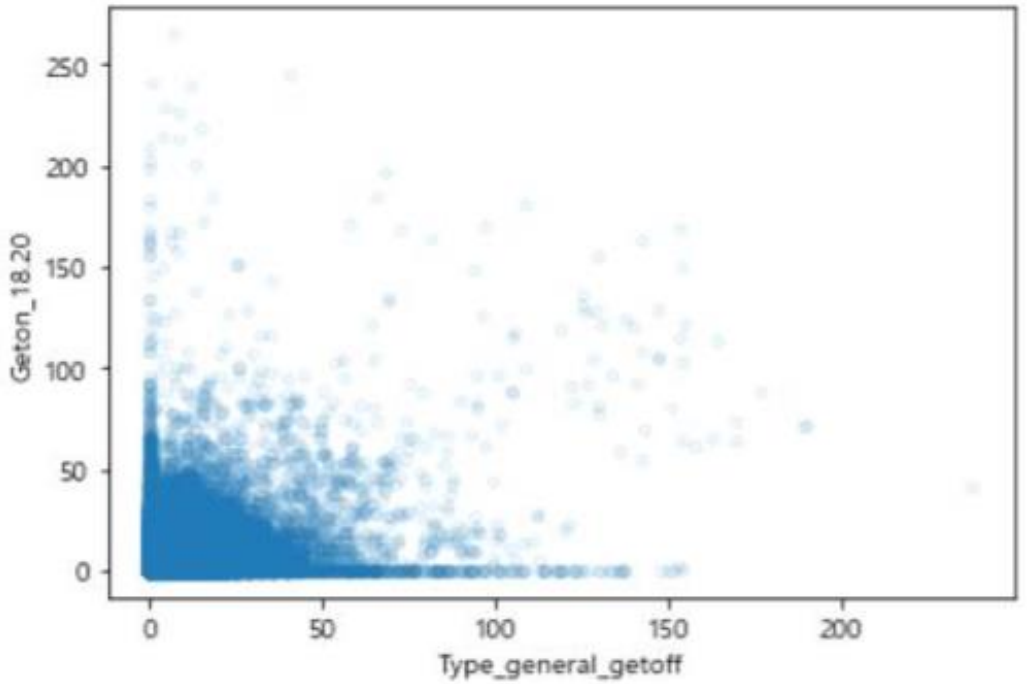
➡ 하차인원의 유형 비율 정보가 없는 경우 general과 others에 동일한 비율로 입력함.

변수 정리 승하차 승객의 유형을 나타내는 변수

일반 승객의 승·하차 인원과 반응변수 관계

Type_general_geton : (일반 + 청소년)



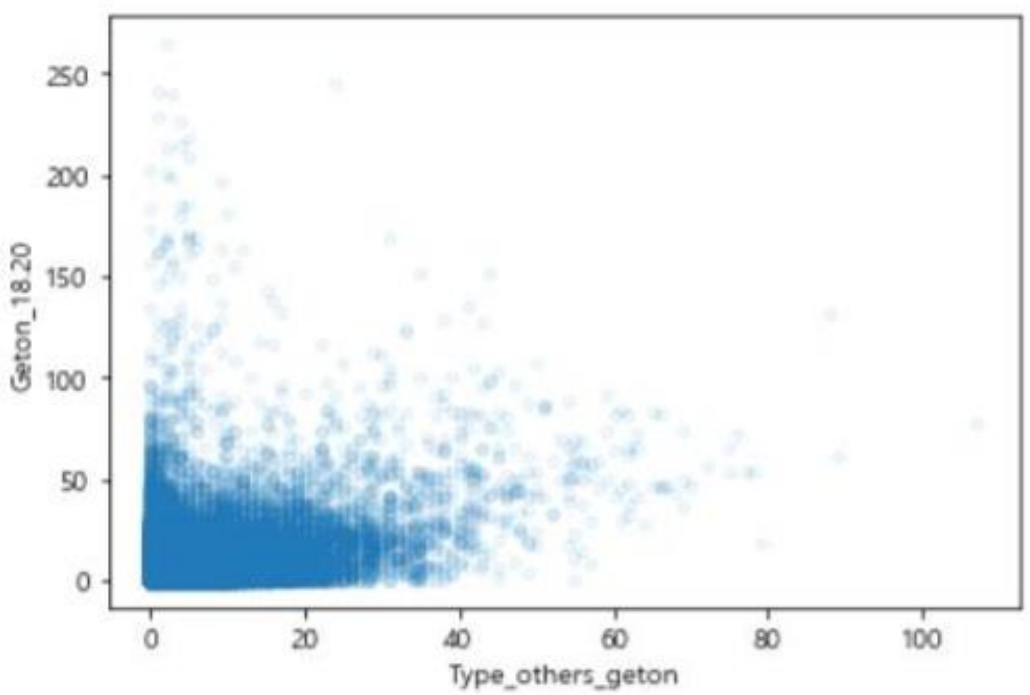


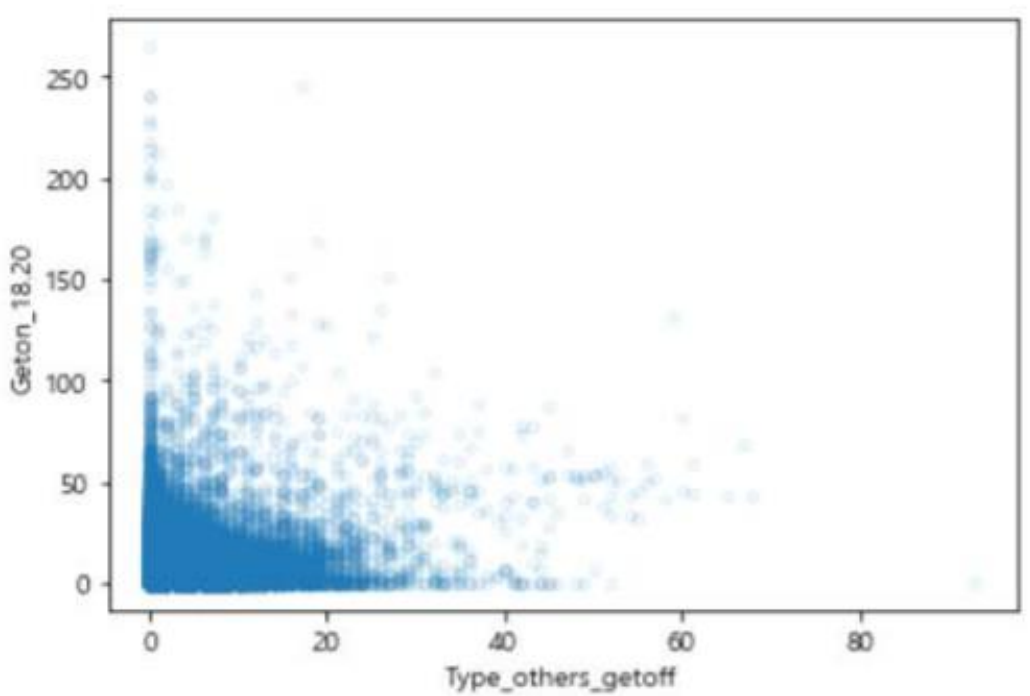
일반 승객의 승·하차 인원이 많을수록 반응변수 값도 커지는 경향

변수 정리

나머지 승객의 승·하차 인원과 반응변수 관계

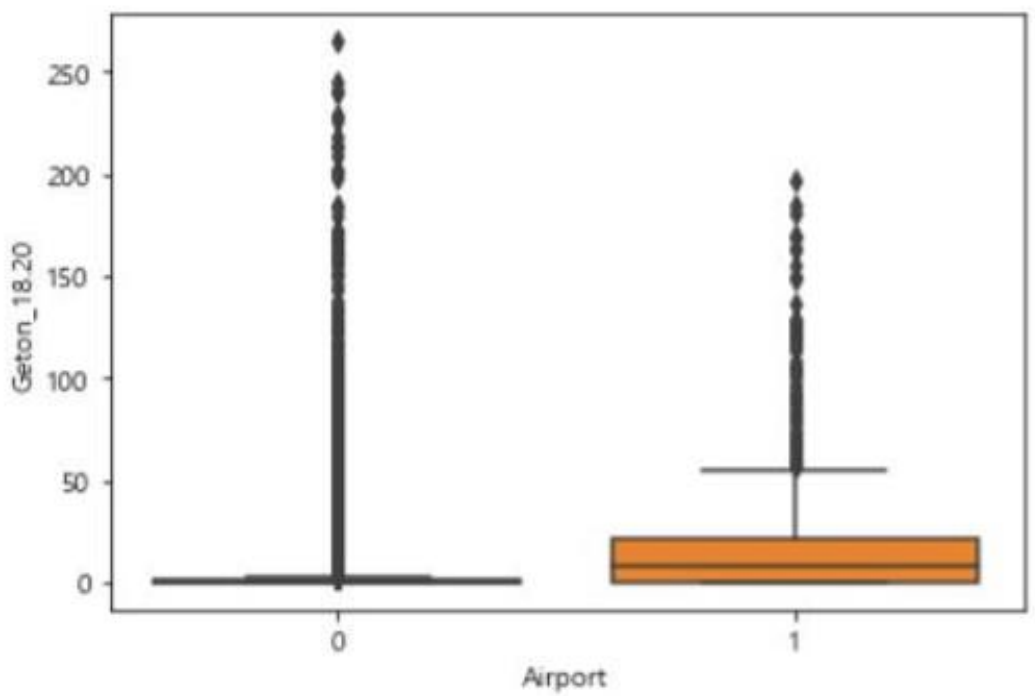
Type_others_geton : (일반 + 청소년) 제외

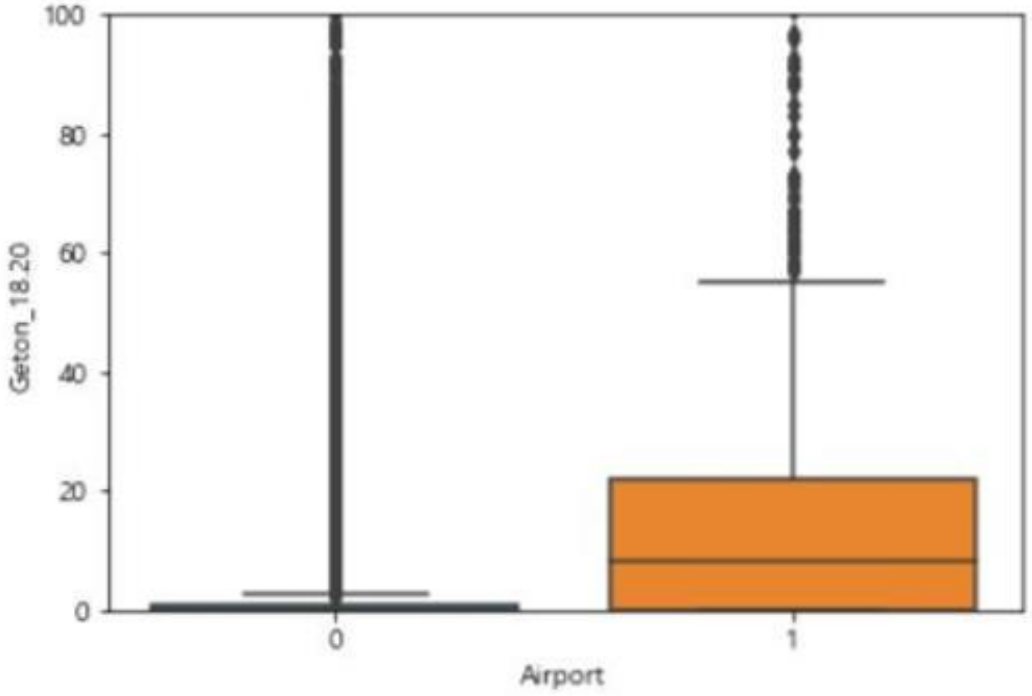




나머지 승객의 승·하차 인원이 많아도 반응변수 값은 거의 일정한 경향

변수 정리 공항 정류장과 반응변수 관계





공항 정류장에서 승차인원이 많음.

변수 생성 : Airport (정류장 이름에 '공항'이 포함되면 1, 아니면 0)

변수 정리 주거지 특성 변수 추가

7월 기준 3개월 연속 제주 땅값이 하락 추세인 가운데, 지역별로 상승률 차이가 뚜렷.

	땅값 상승 (14개 지역)	땅값 하락 (12개 지역)
읍면동	봉개동, 용강동, 회천동, 월평동, 영평동, 성산읍, 구좌읍, 아라동, 화북동, 애월읍, 표선면, 남원읍, 안덕면, 우도면	삼도동, 용담동, 일도동, 건입동, 색달동, 상예동, 하예동, 중문동, 회수동, 대포동, 하원동, 서귀동
특징	성산읍에 예정된 제주 제2공항 개발에 따른 기대수요와 경제적 파급효과 -> 동부지역 땅값 상승세	인구가 이탈되는 공동화 현상이 심화되고 있는 원도심 지역

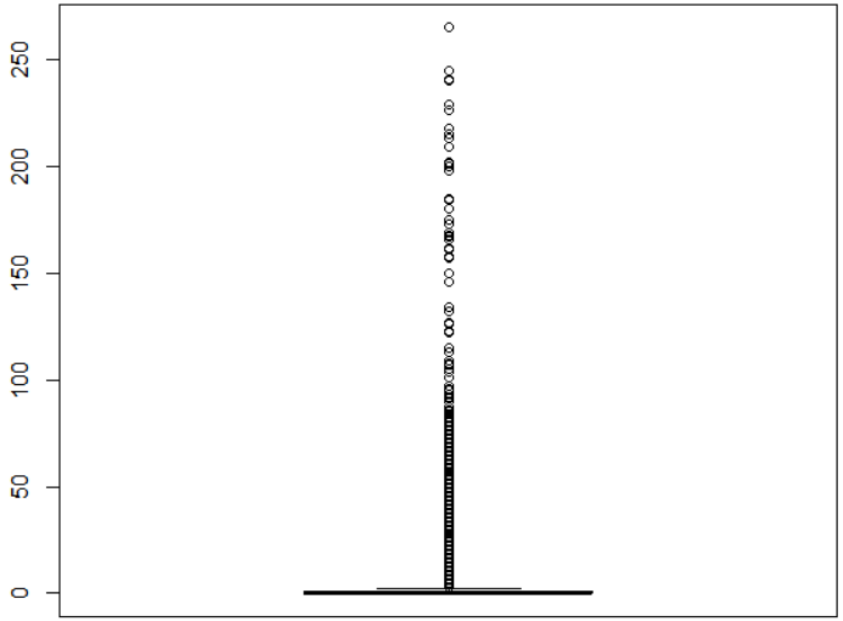
※ 출처 : 제주신보 2019.9.9 <제주 땅값 하락세 속 지역차 뚜렷>

▷ Price_up, Price_down 변수 생성하여 정류장을 범주화함.

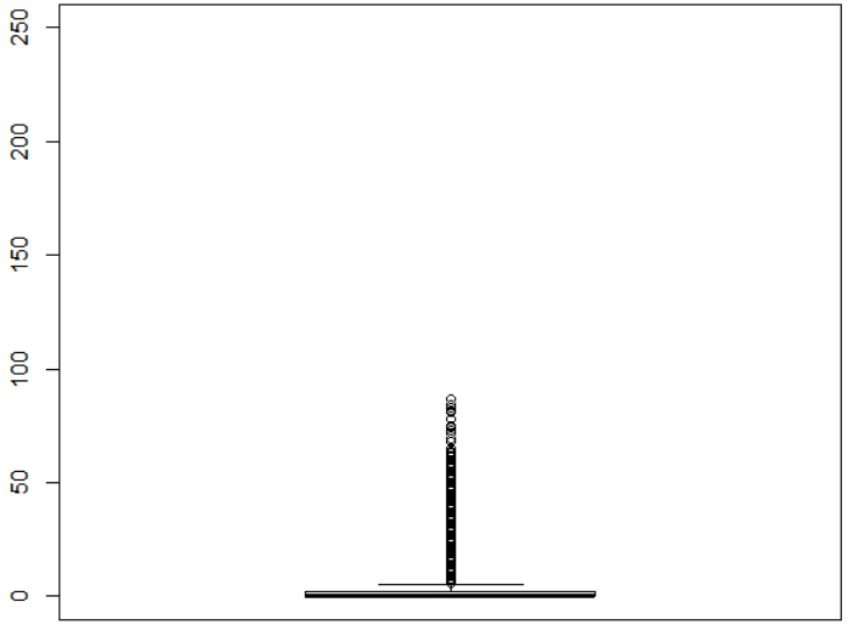
총 58개 지역중 7월, 8월 연속적으로 각각 땅값이 상승하고 하락한 지역.
땅값이 상승하는 지역에 위치한 정류소는 주거지 인근 정류소 특징이 나타난다고 판단

변수 정리 주거지 특성 변수 추가

땅값 상승 지역 정류장의 퇴근시간 승차량



땅값 하락 지역 정류장의 퇴근시간 승차량

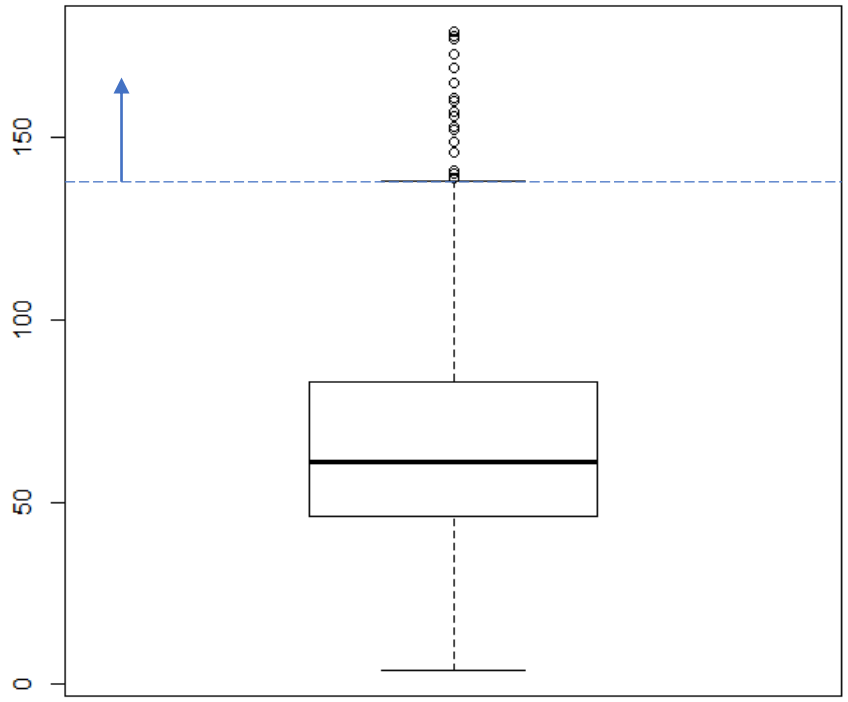


▷ Price_up, Price_down 변수 생성하여 정류장을 범주화함.

총 58개 지역중 7월, 8월 연속적으로 각각 땅값이 상승하고 하락한 지역.
땅값이 상승하는 지역에 위치한 정류소는 주거지 인근 정류소 특징이 나타난다고 판단

이상치 처리

▷ Bus_interval (배차 간격)



<배차 간격 boxplot>

만약 버스가 정류장에 멈춰도 승차한 승객이 없다면

배차 간격이 실제보다 크게 계산될 수 있음.

값이 Q3+1.5*IQR 이상이면

Q3+1.5*IQR 값으로 대체

최종 변수

관측치 (obs.) : 286,706 개

반응 변수

- Geton_18.20

노선 변수

- Route_score
- Bus_interval

날씨 변수

- Typhoon
- prcp

날짜 변수

- Holiday
- Mon
- Tue
- Wed
- Thu
- Fri
- Sat
- Sun

정류장 변수

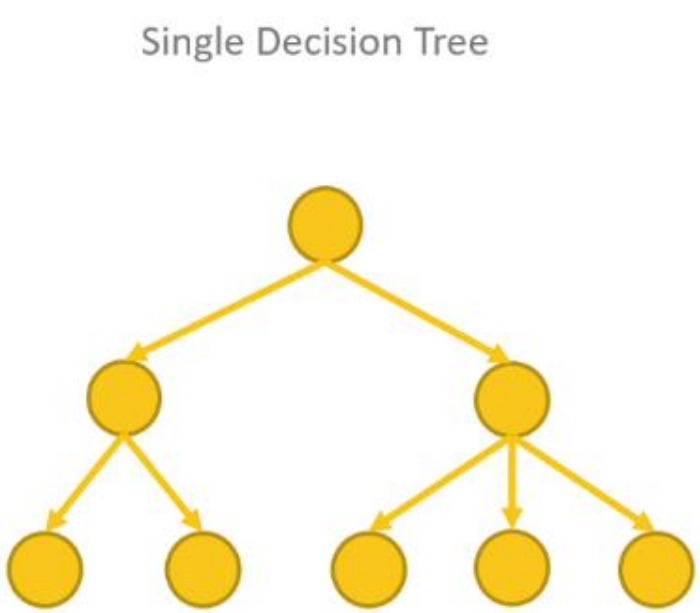
- Price_up
- Price_down
- Airport
- Station_g2
- Station_g3

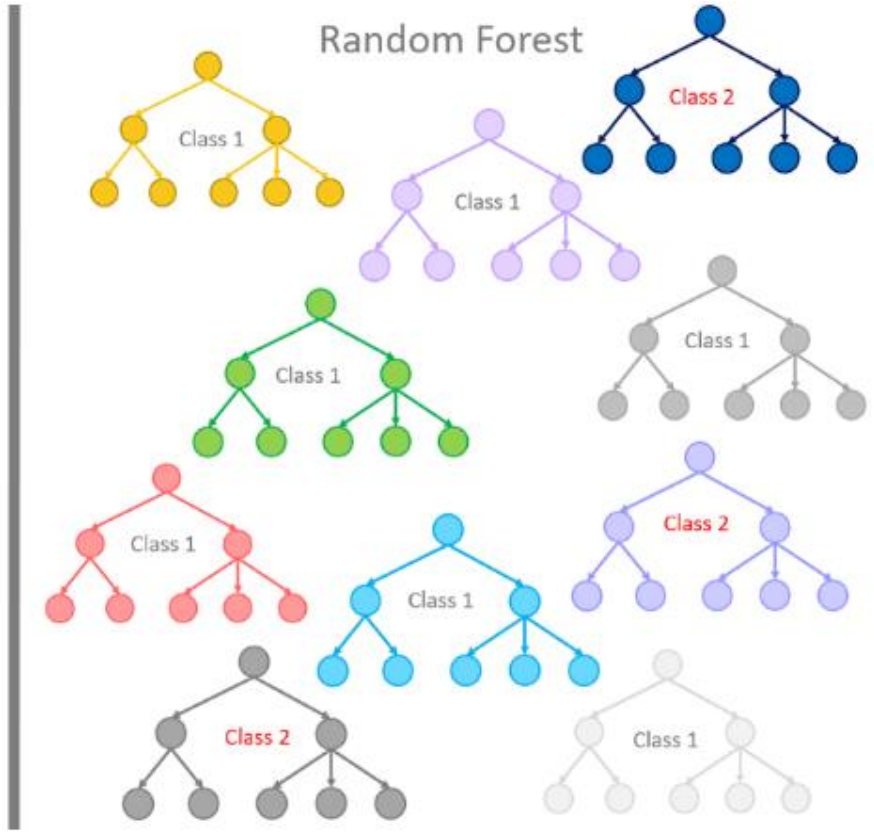
승하차 인원 변수

- Geton_6.9
- Geton_9.12
- Getoff_6.9
- Getoff_9.12
- Geton_total
- Getoff_total
- Commuter_count
- Type_general_geton
- Type_others_geton
- Type_general_getoff
- Type_others_getoff

모델링

랜덤포레스트





- 앙상블 기법의 대표 모델
- 결정트리와 bagging 결합 모델
- 과적합을 방지하고 안정성 높음
- 스케일에 구애받지 않음
- 다중공선성 영향이 적음

교차검증을 사용한 그리드서치

GridSearch

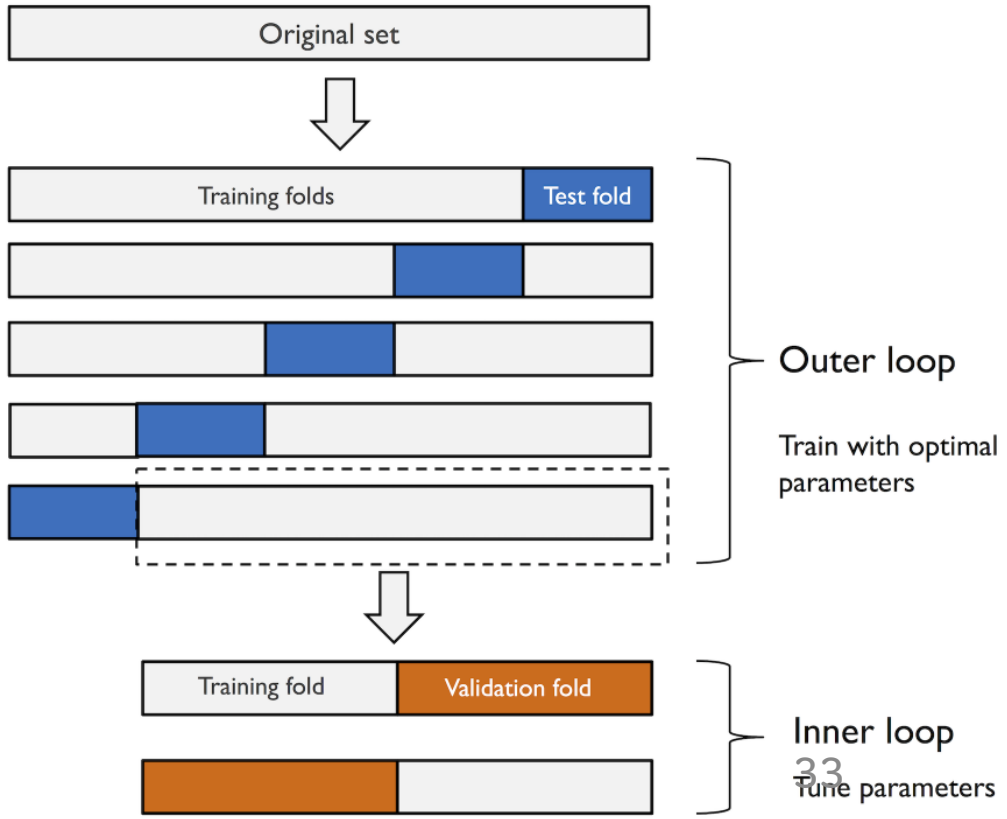
관심있는 매개변수들로 가능한 모든 조합을 시도하여 최적의 매개변수를 찾는 방법

➡ 매개변수를 튜닝하여 일반화 성능 개선

CrossValidation

vaildatation set으로 여러 번 모델 평가를 진행하는 방법

➡ 데이터 편중 방지 & 더욱 일반화된 모델 생성



최종 모델

RandomForestRegressor(n_estimators = 1000

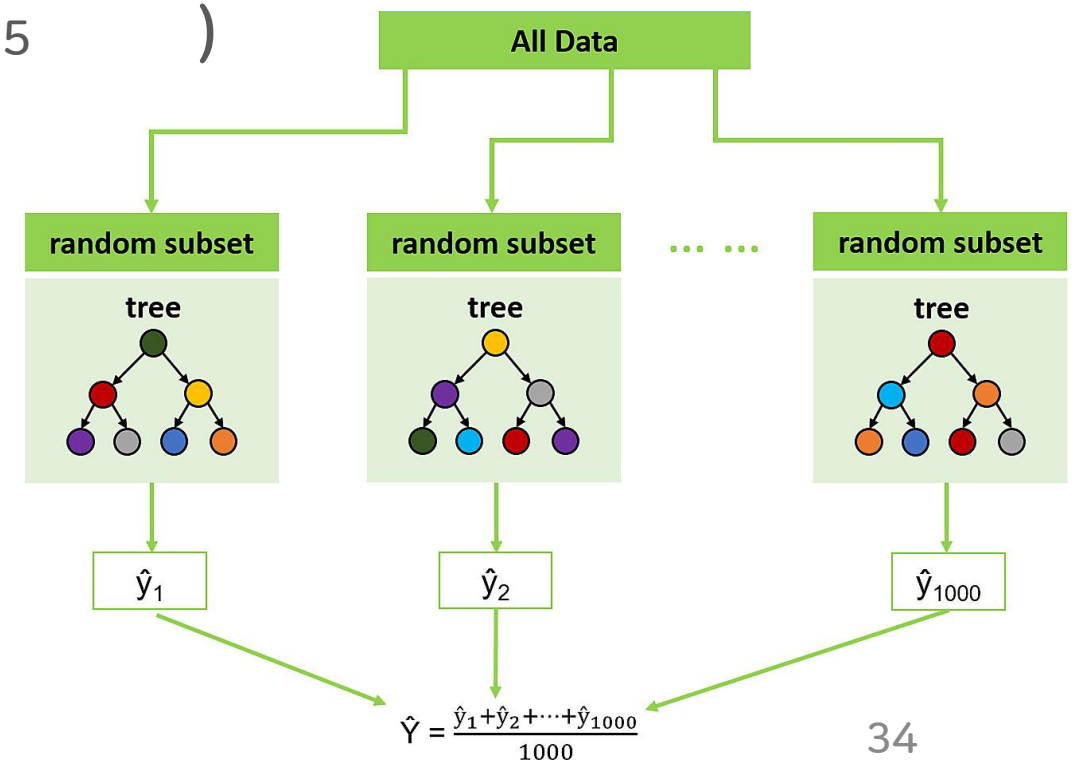
max_features = 32

min_sample_splits = 5)

➡ 최대 32개의 특성을 선택하여

하나의 노드에 최소 5개의 샘플이 남을 때까지

가지치기를 진행한 회귀 트리 1000개 생성



평가 및 결론

최종 모델 성능 평가

최종 모델 정확도 72%

Mean absolute error = 1.25

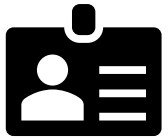
Mean squared error = 8.79

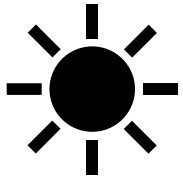
Root mean squared error = 2.96

변수 중요도 평가

퇴근시간 승차인원이 많아지는 요인


출근시간대 이용량이 많은 노선의 정류소


통근자가 많이 이용하는 노선의 정류소


오전 승차량이 많은 노선의 정류소


단거리 운행 노선


배차간격이 30분 이내인 노선의 정류소


일반, 청소년 승객 비율이 높은 노선의 정류소

예측 모델 활용 방안

▷ 주차난 해소

제주신보 2019.09.23 <서귀포시 중앙동·대정 상모리 구간 '일방통행' 지정>

서귀포시에서는 양방통행 구간에서 도로불법 주차로 인해 교통혼잡이 심화되자

최근 3년동안 12개 구간을 양방통행 -> 일방통행으로 지정

승차인원 예측을 통해 교통 혼잡 구역을 찾아내고 인근 지역 주차 문제를 해소

예) 공영주차장 등 시설 마련

예측 모델 활용 방안

▷ 카셰어링 사업 추진

대중교통 접근성이 떨어지는 지역에 노선을 늘리지 않고 단거리 카셰어링을 도입하는 방안

예) 제주 첨단 과학 기술 단지

- 대중교통 접근성이 떨어져 대다수의 직원들이 자가용으로 출퇴근*
- 라스트 마일 모빌리티 : 대중교통역 근처에 단거리 카셰어링을 도입하는 방식

*출처 : 제주신보 2019.09.02 <'라스트 마일 모빌리티' 초소형 전기차 활용 주목>

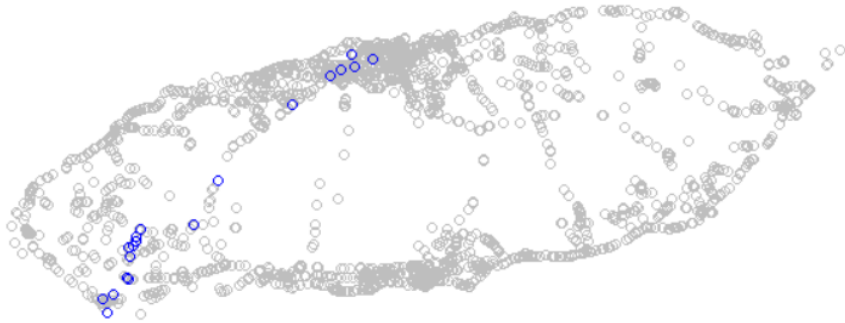
예측 모델 활용 방안

▷ Station_g3 = 1 정류소의 교통정체

- 시간대별 승차량이 골고루 많았던 Station_g3 정류소

제주버스터미널, 제주대학교, 동문로터리(동문시장),중앙로(국민은행) ,
제주도청신제주로터리 , 한라병원 ,노형오거리 제주국제공항(구제주방면) ,
제주국제공항(신제주방면) ,중앙로터리(동),중앙로터리 , 제주시청(아라방면),…

- 정류소별 노선 개수 순위에서도 상위권에 속함.



<Station_g3=1 정류소에서 고속도로로 이어지는 노선 예시>

➡ 경유(환승) 정류장 특징 ?

경유 정류장의 이용량은 (서귀포로 이어지는) 평화로 등 고속도로 정체와 연결됨.

따라서 공항과 경유정류장 인근의 교통량을 분산시키기 위해 우회도로나 지하차도를 개설하는 방안


감 사 합 니 다