2조

일별 지하철 이용자 수에 대한 시계열 분석

김민규 노현경 유태윤

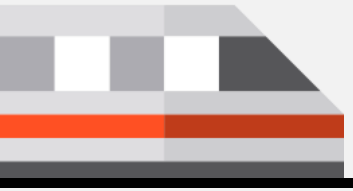


Table of CONTENTS

01 Introduction

- 출처 : 서울열린데이터광장
- 모든 역의 2015~2019년 일별 승하차 승객수

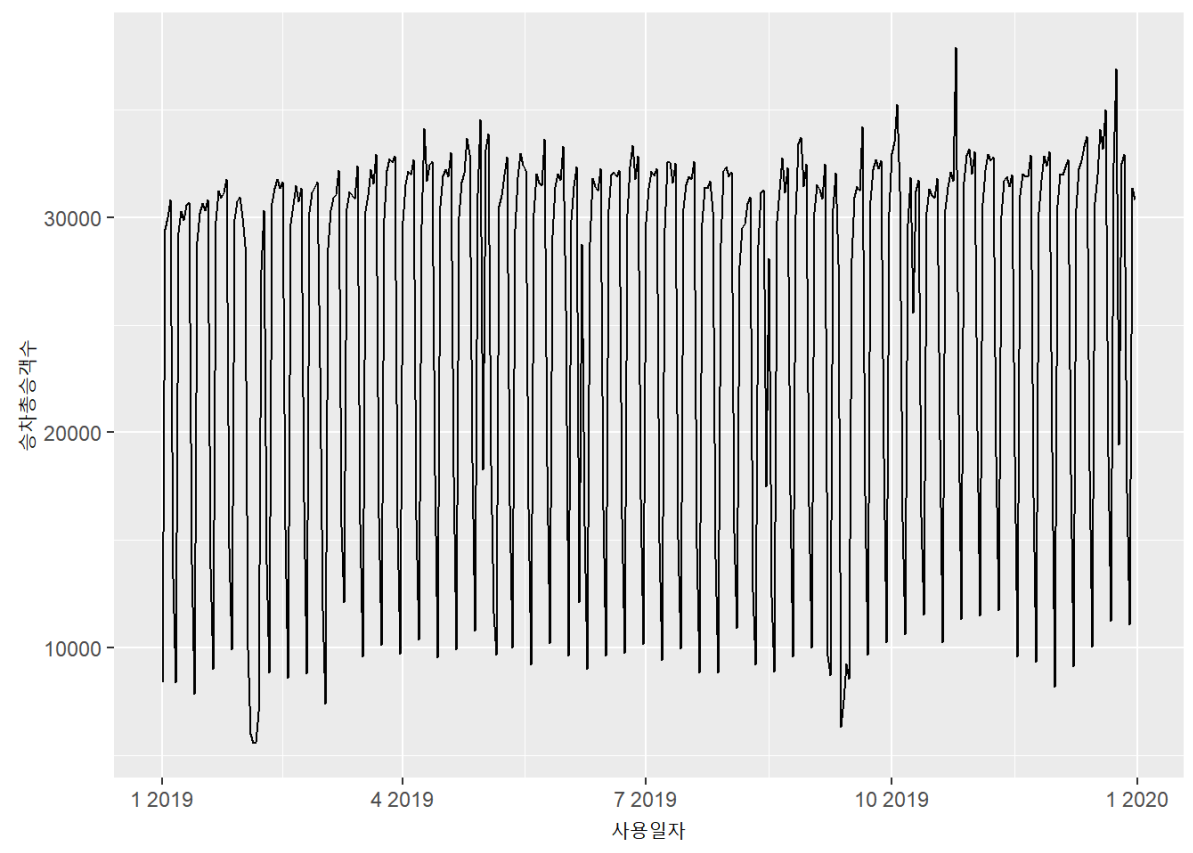
	A	B	C	D	E	F	G
1	사용일자	노선명	역ID	역명	승차총승ㄱ	하차총승ㄱ	등록일자
2	20170101	경원선	1907	가능	4194	3852	20170109
3	20170101	경원선	1908	녹양	2612	2468	20170109
4	20170101	경원선	1909	양주	5036	5004	20170109
5	20170101	경원선	1910	덕계	1145	1123	20170109
6	20170101	경원선	1911	덕정	3823	4188	20170109

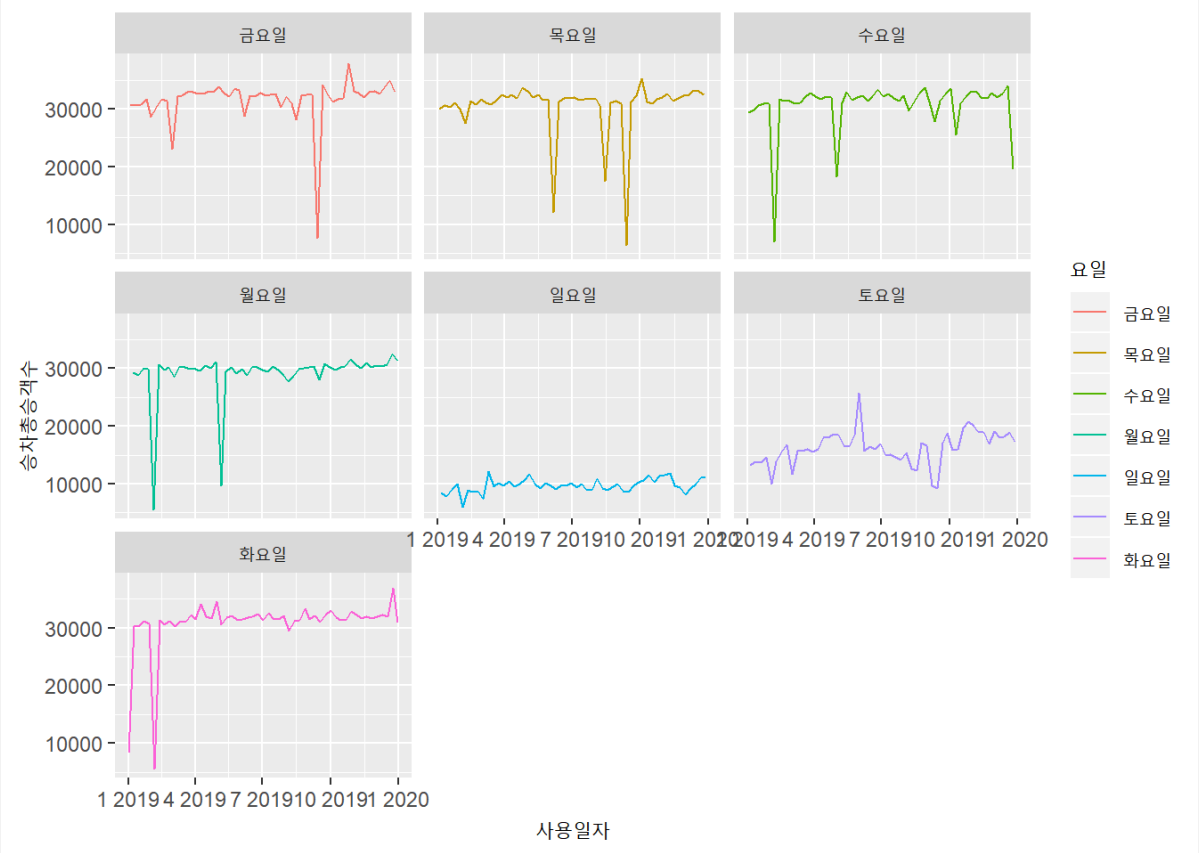
01 Introduction

- 2호선의 승차 인원을 중심으로 분석
- 특성에 따라 역을 그룹으로 나누기
  - ex) 주말에 이용자가 많은 역, 평일에 이용자가 많은 역
- 각 그룹에서 대표역 1개 선정
- 각 역에 대해 18년까지의 데이터로 모델 적합 후 19년 데이터로 validation

02 Grouping

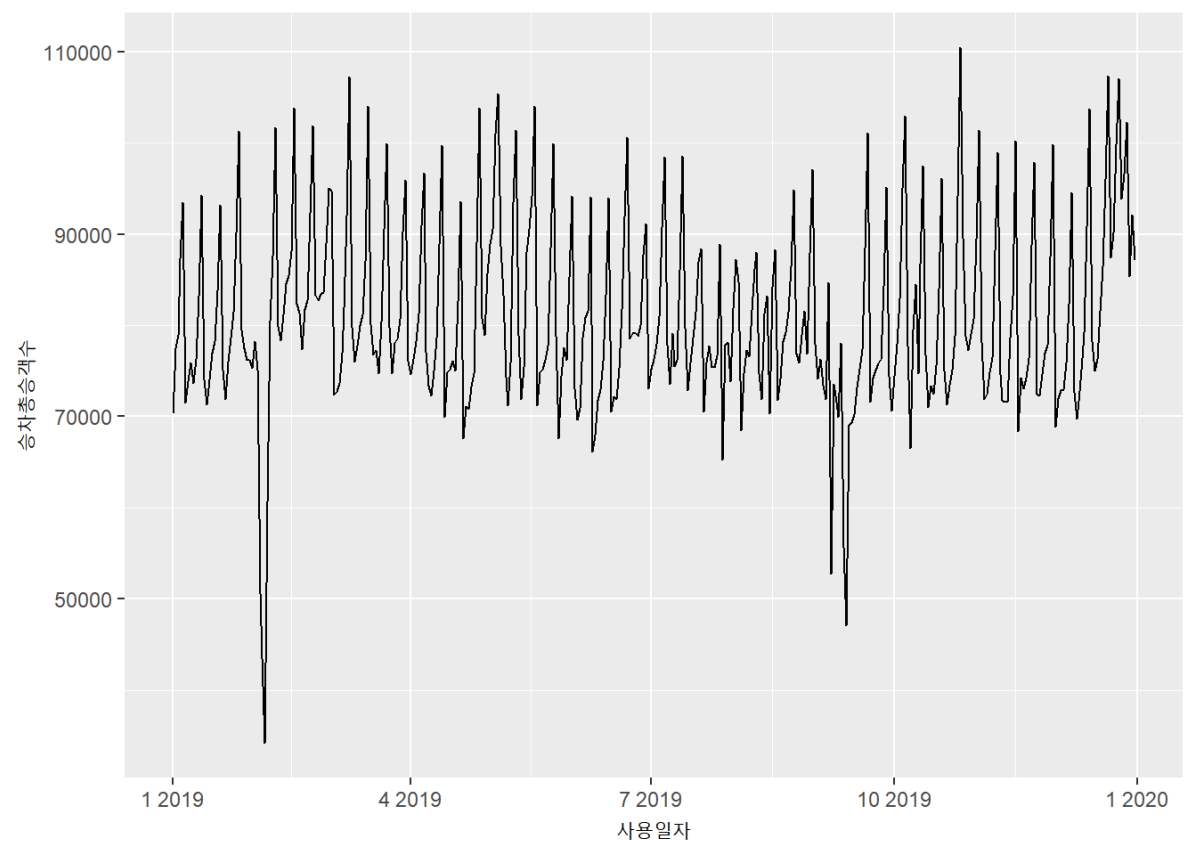
Group 1 : 승차인원이 평일에 많고 주말에 적은 group
대표역 : 시청

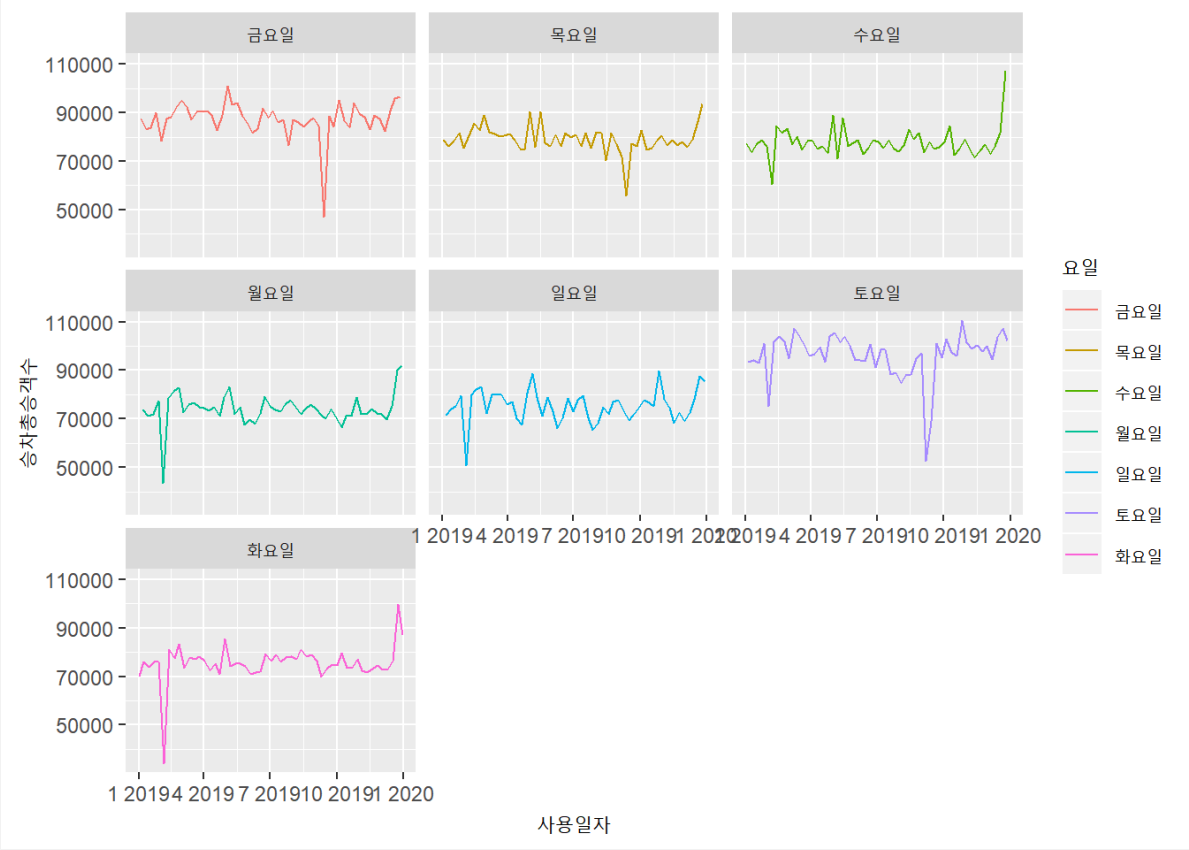




02 Grouping

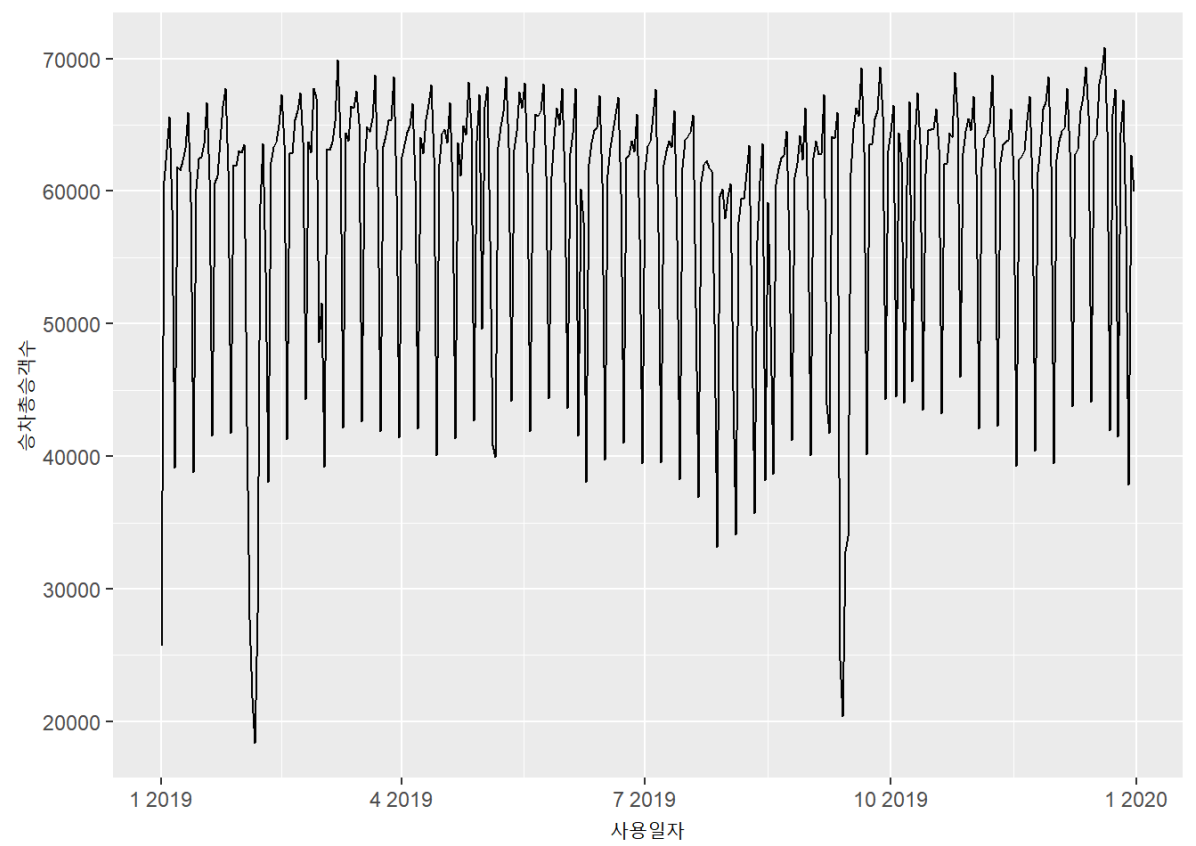
Group 2 : 승차인원이 평일보다 금,토에 많은 group
대표역 : 홍대입구

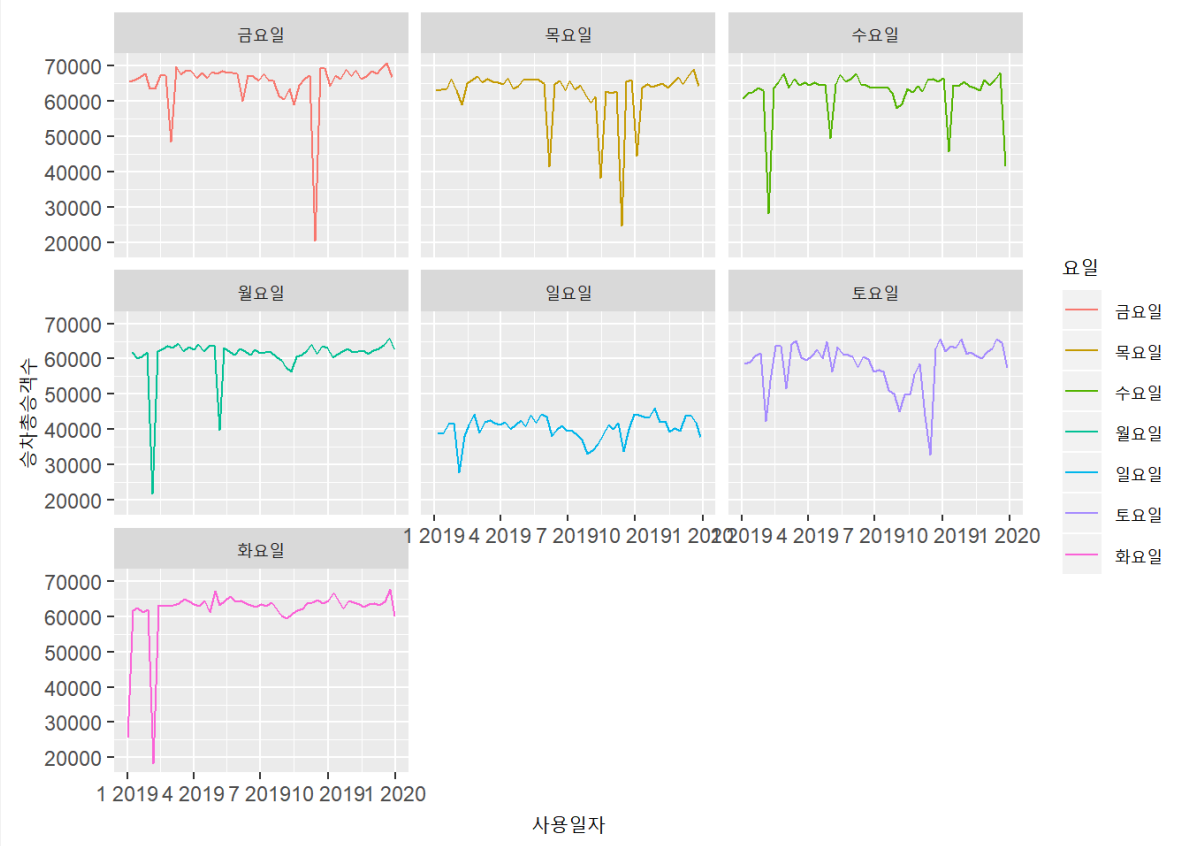




02 Grouping

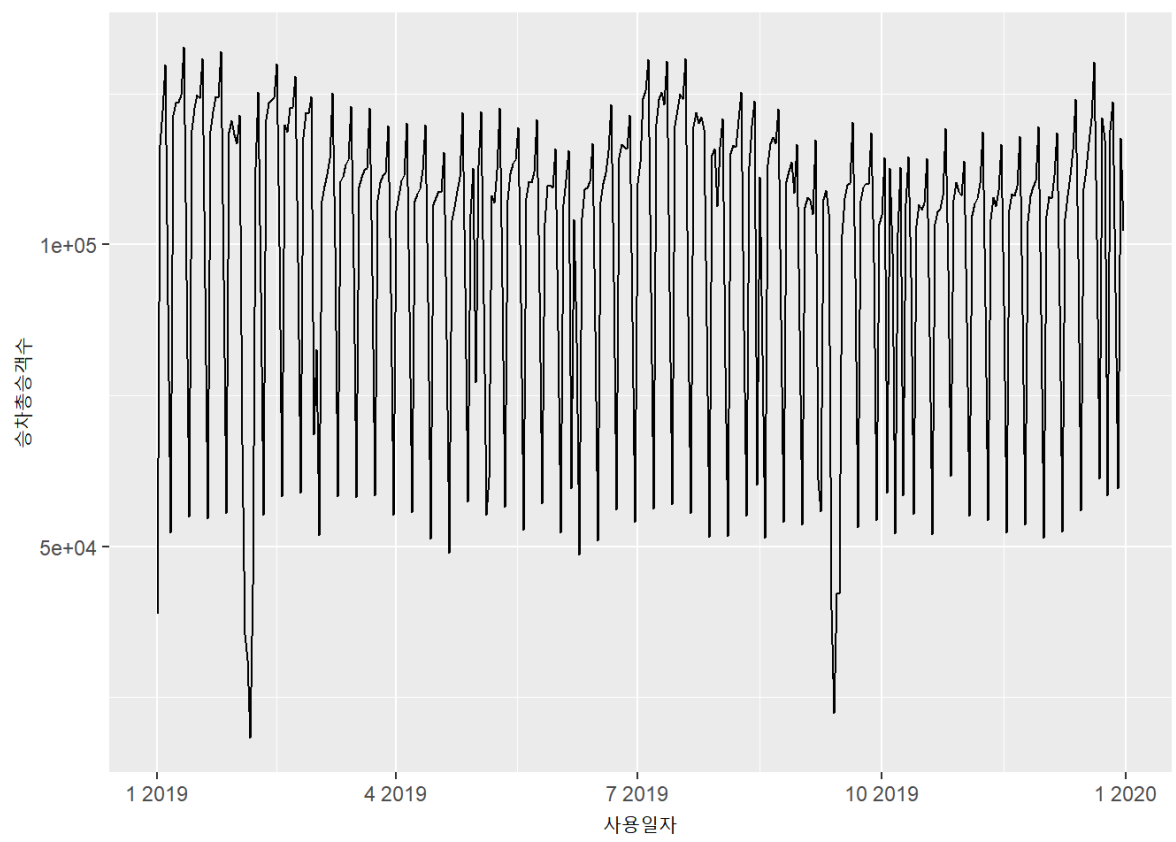
Group 3 : 일요일만 승차인원이 적은 group
대표역 : 신도림

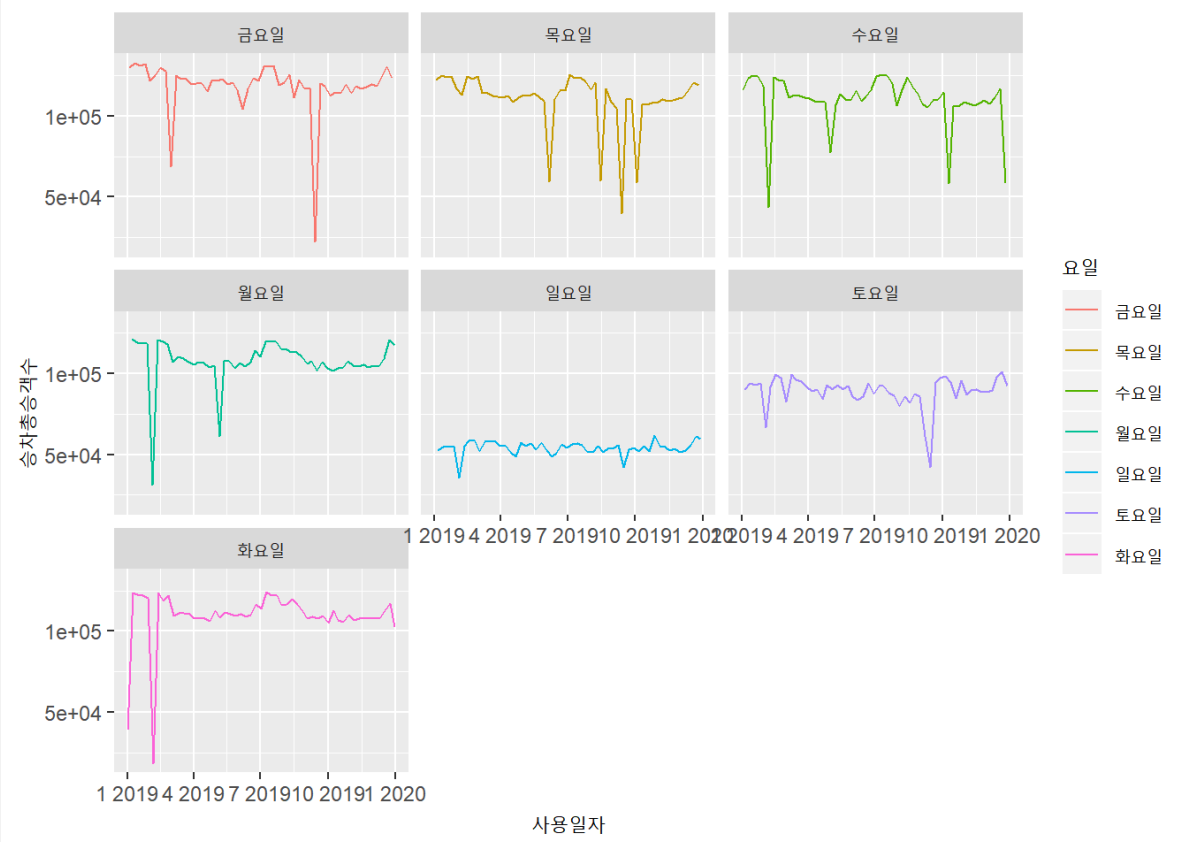




02 Grouping

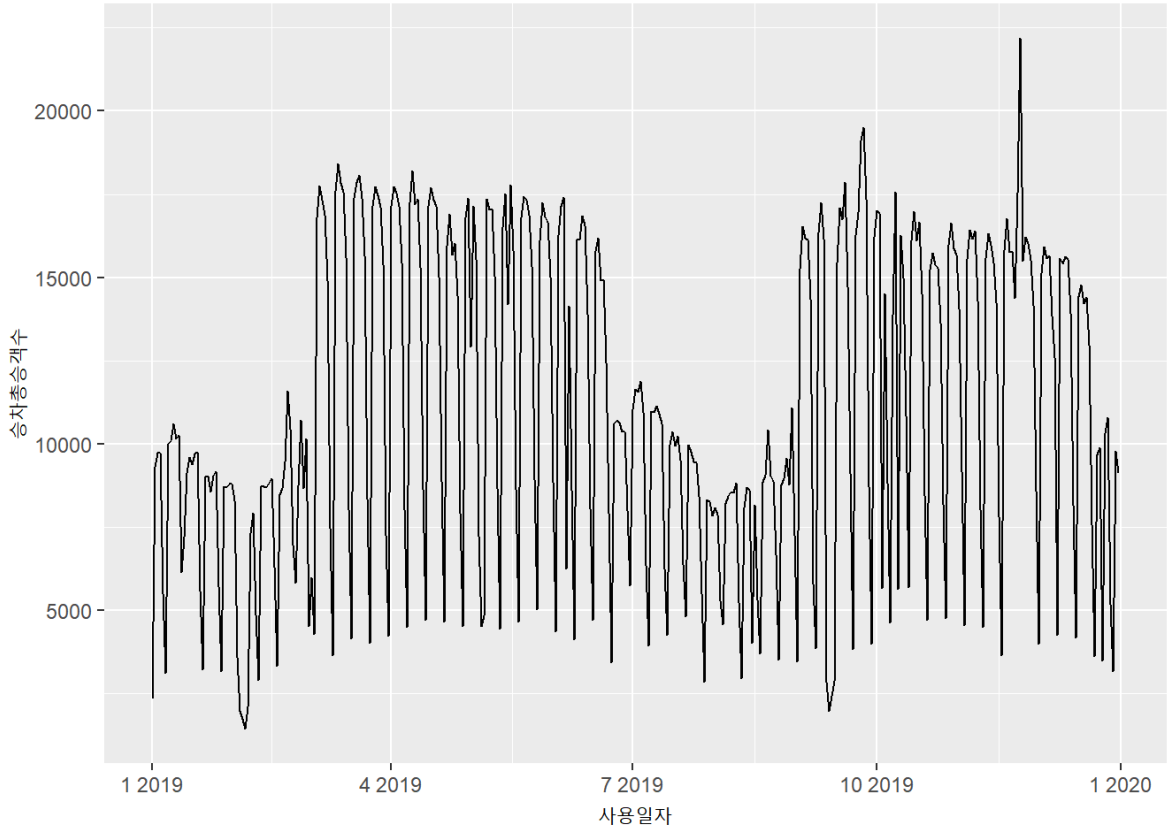
Group 4 : 일요일의 승차인원이 가장 적고 토요일이 중간 정도인 group
대표역 : 강남

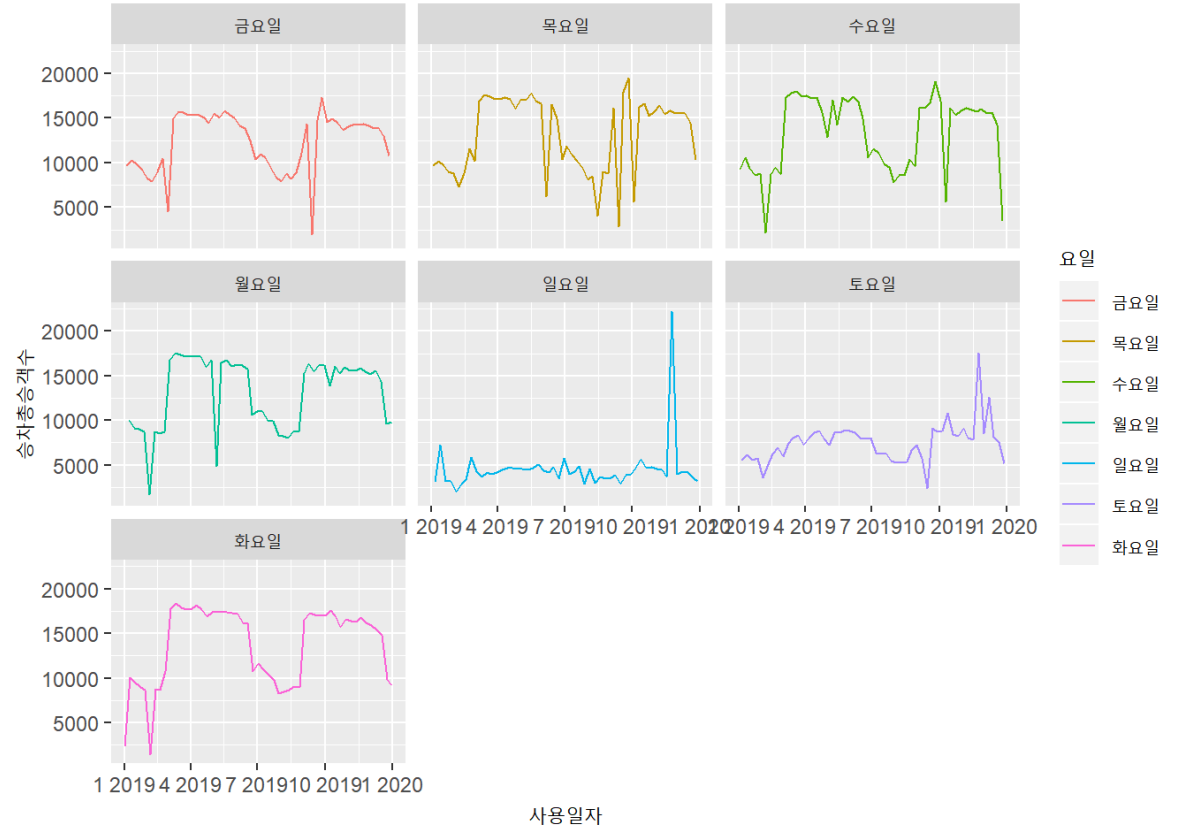




02 Grouping

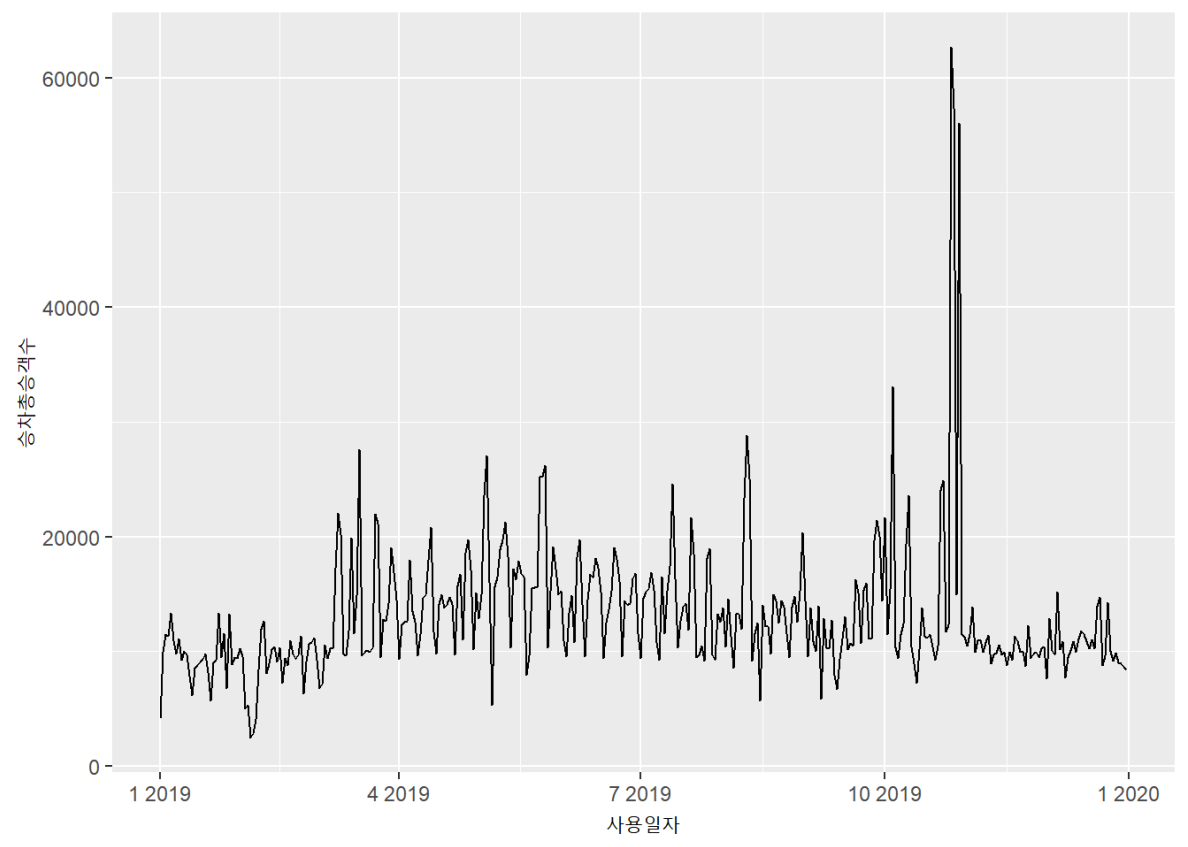
Group 5 : 특정 기간에만 승차인원이 많은 group
대표역 : 한양대입구(학기 중과 방학 때)

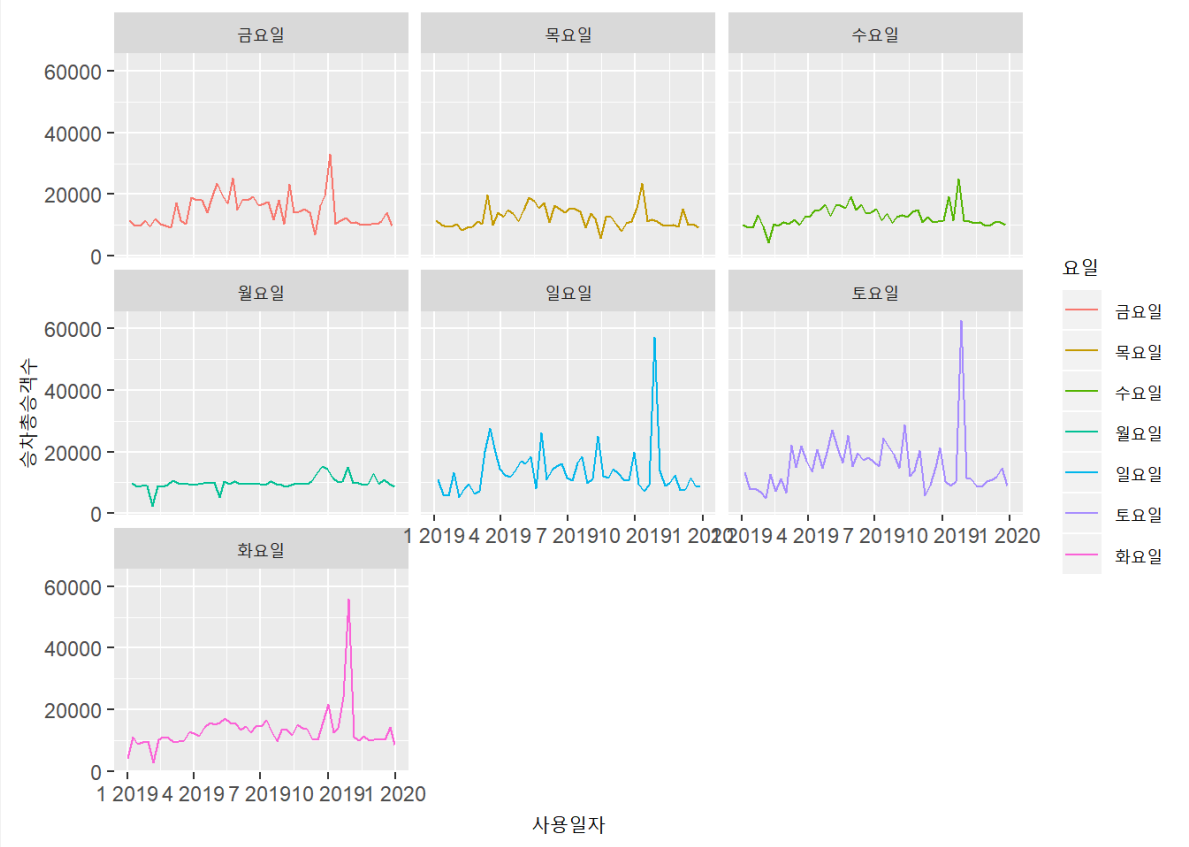




02 Grouping

Group 6 : 특정 날에만 승차인원이 많은 group
대표역 : 종합운동장(19년 10월 26일,27일,29일 방탄소년단 콘서트)

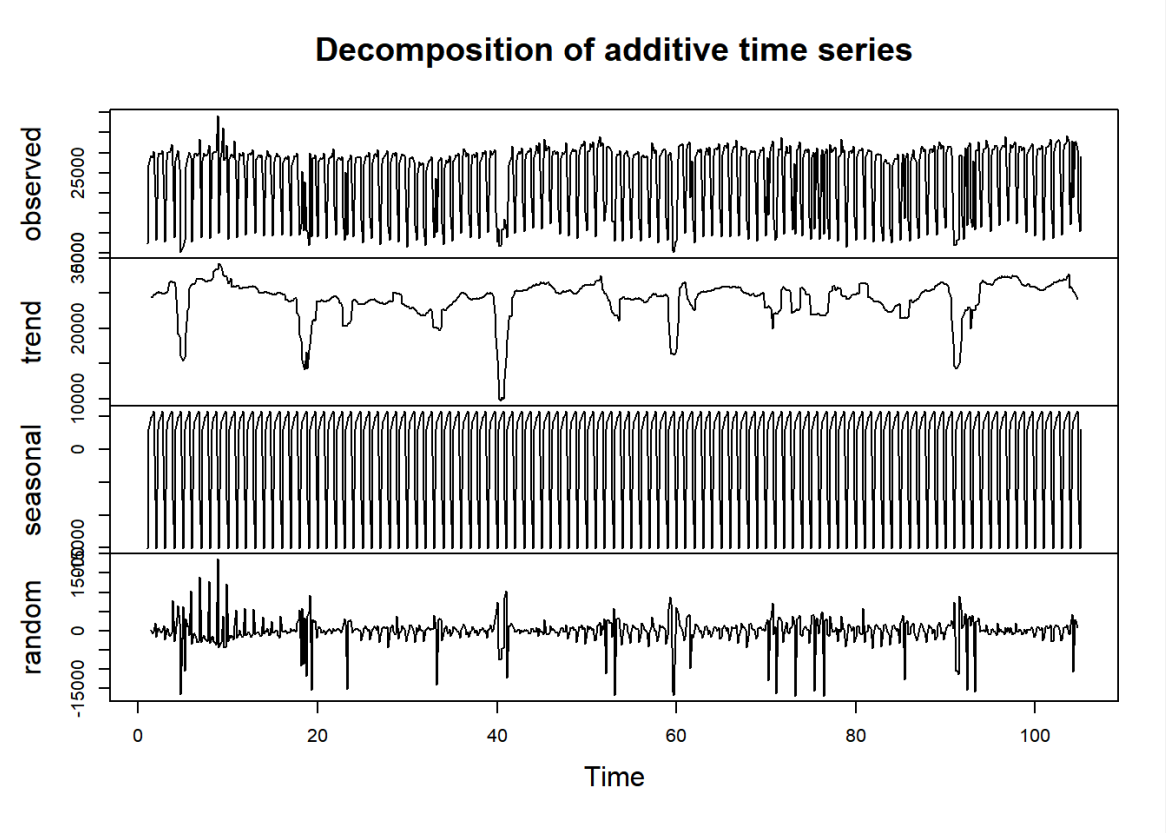




- 그 밖에도 크고 작은 행사들이 있어 불규칙한 데이터를 보여준다.

03 모델 선택 및 예측

Group 1 (시청) : 승차인원이 평일에 많고 주말에 적은 group
대표역 : 시청

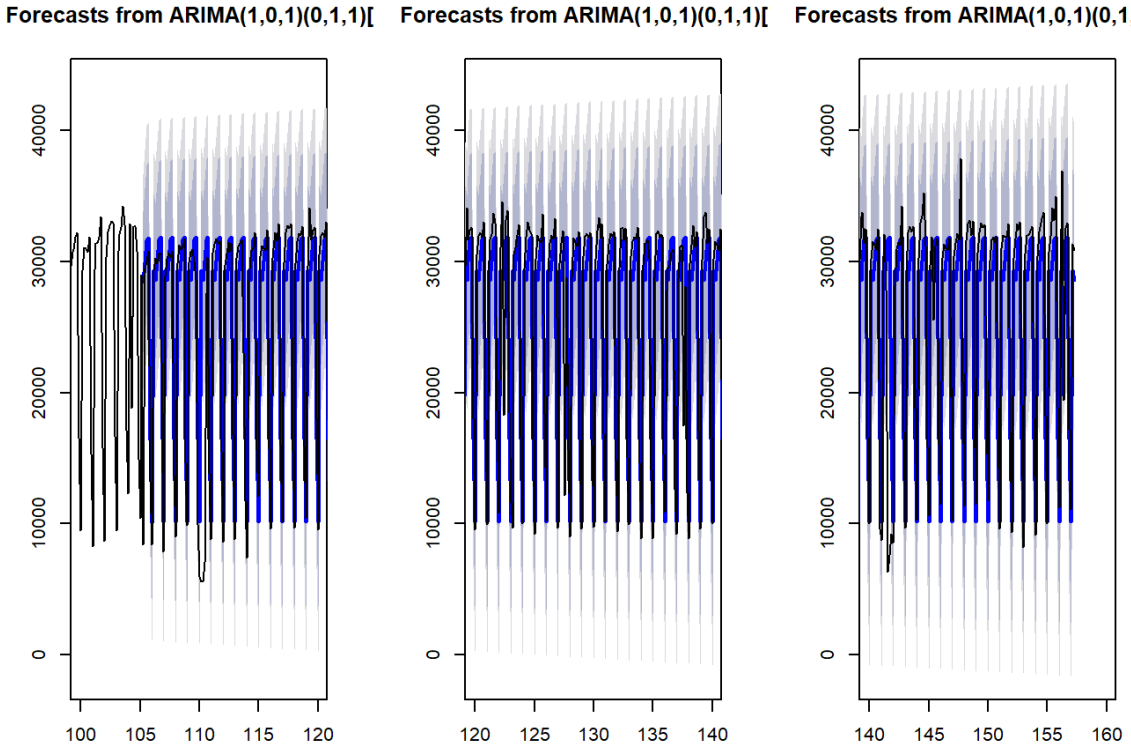


- 7일(1주)를 주기로 하였을 때,
  seasonal 성분과 아닌 성분의 분리가 잘 이루어진다.

- auto.arima로 ARIMA(1,0,1)(0,1,1)[7] 모델 적합

03 모델 선택 및 예측

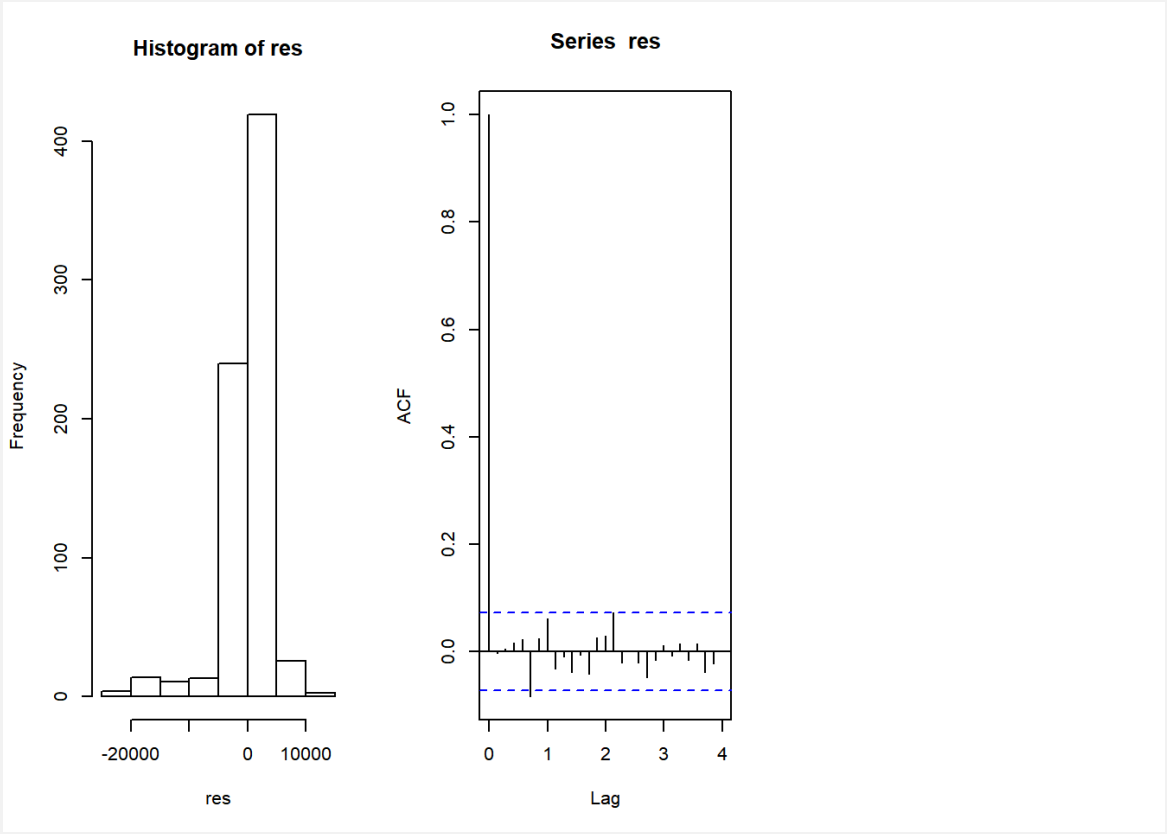
Group 1 (시청) : 승차인원이 평일에 많고 주말에 적은 group
대표역 : 시청



- 17~18년 데이터로 19년 데이터에 대한 예측
- 예측값의 분산이 많이 커지지 않지만, 예측을 잘 하지 못한다.

03 모델 선택 및 예측

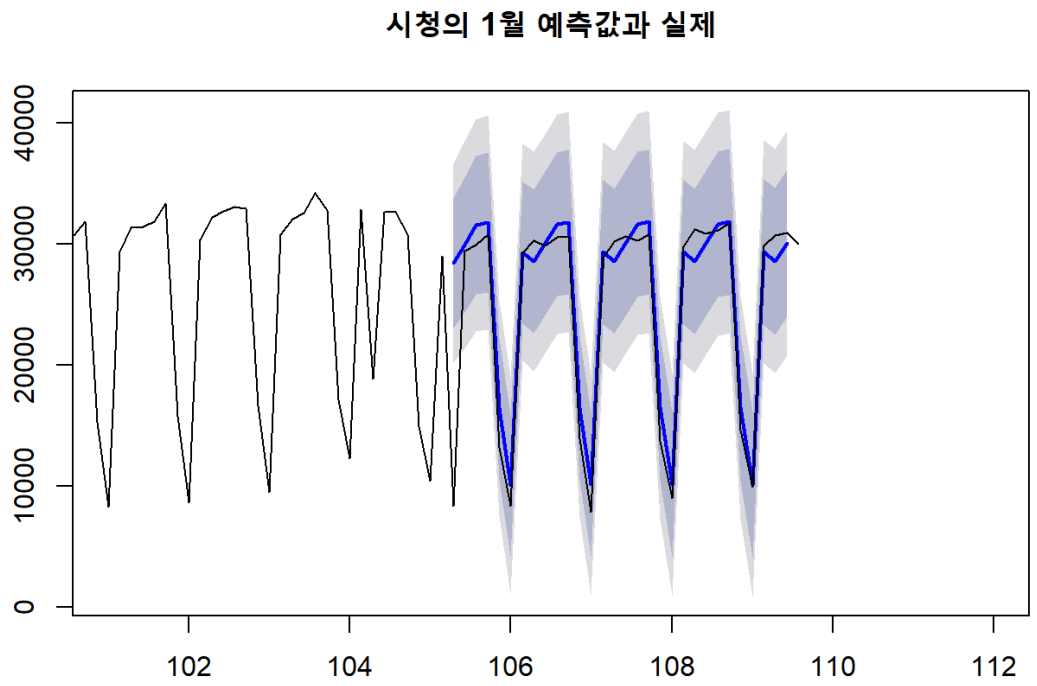
Group 1 (시청) : 승차인원이 평일에 많고 주말에 적은 group

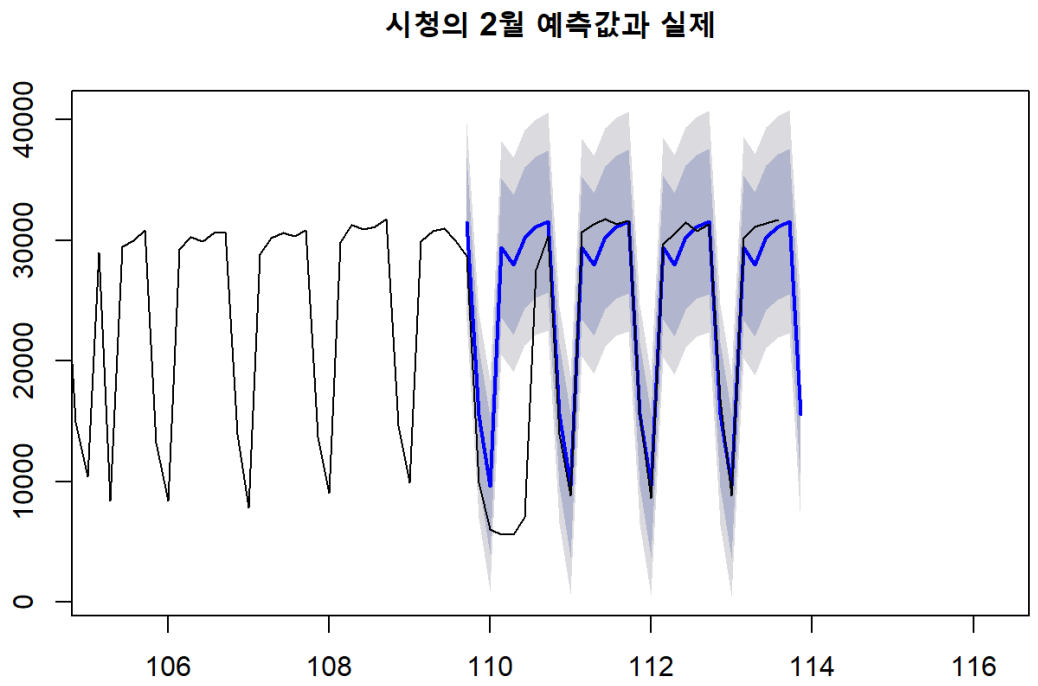


- 잔차의 acf 가 0에 가까운 값이 나온다.
- Box_Ljung test에서도 잔차의 자기상관성이 없다.
- 모델 자체는 유효함을 알 수 있다.
- 예측 범위를 줄여 월별 단위 예측을 시도하였다.
- 모델은 동일하되, 예측 월 직전까지의 데이터로, 그 다음 월을 예측한다.

03 모델 선택 및 예측

Group 1 (시청) : 승차인원이 평일에 많고 주말에 적은 group

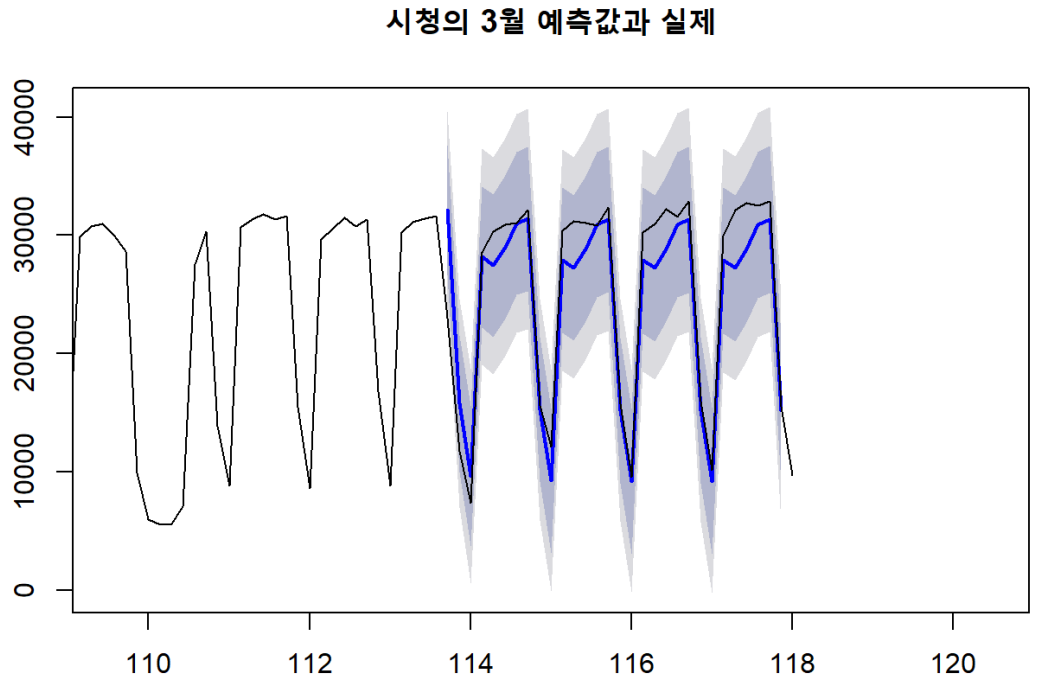


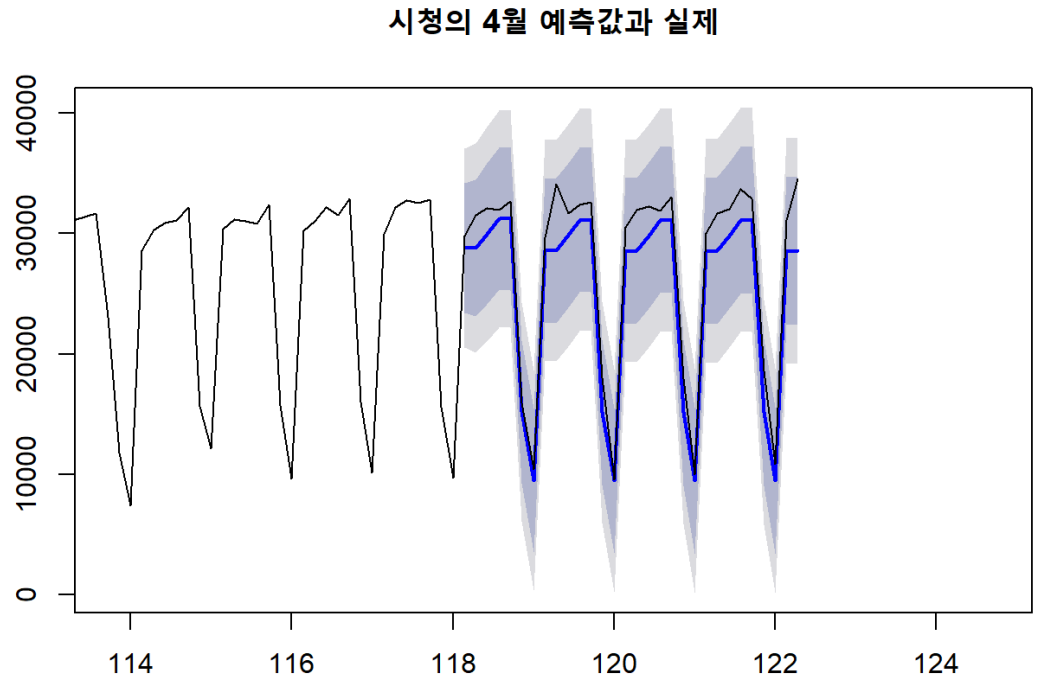


- 2월의 예측하기 어려운 이벤트(설날)을 제외하고는 예측이 양호하다

03 모델 선택 및 예측

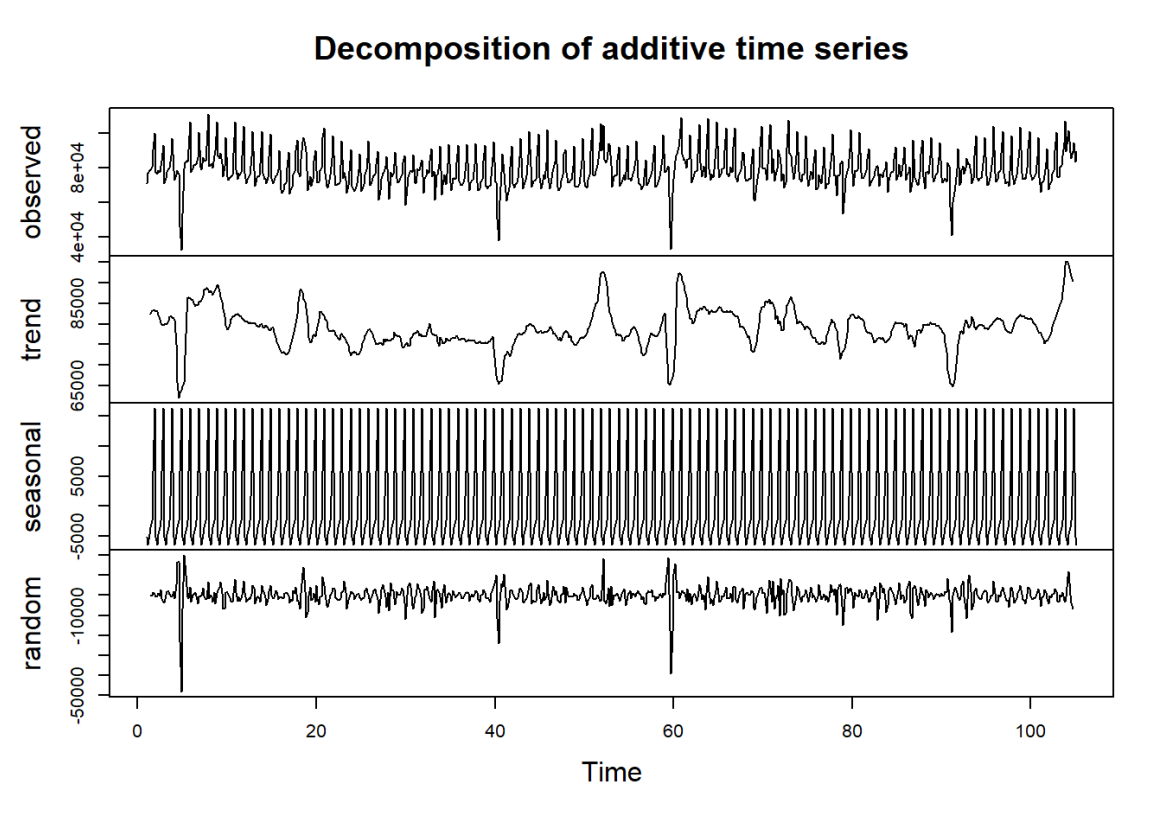
Group 1 (시청) : 승차인원이 평일에 많고 주말에 적은 group





03 모델 선택 및 예측

Group 2(홍대입구) : 승차인원이 평일보다 금,토에 많은 group

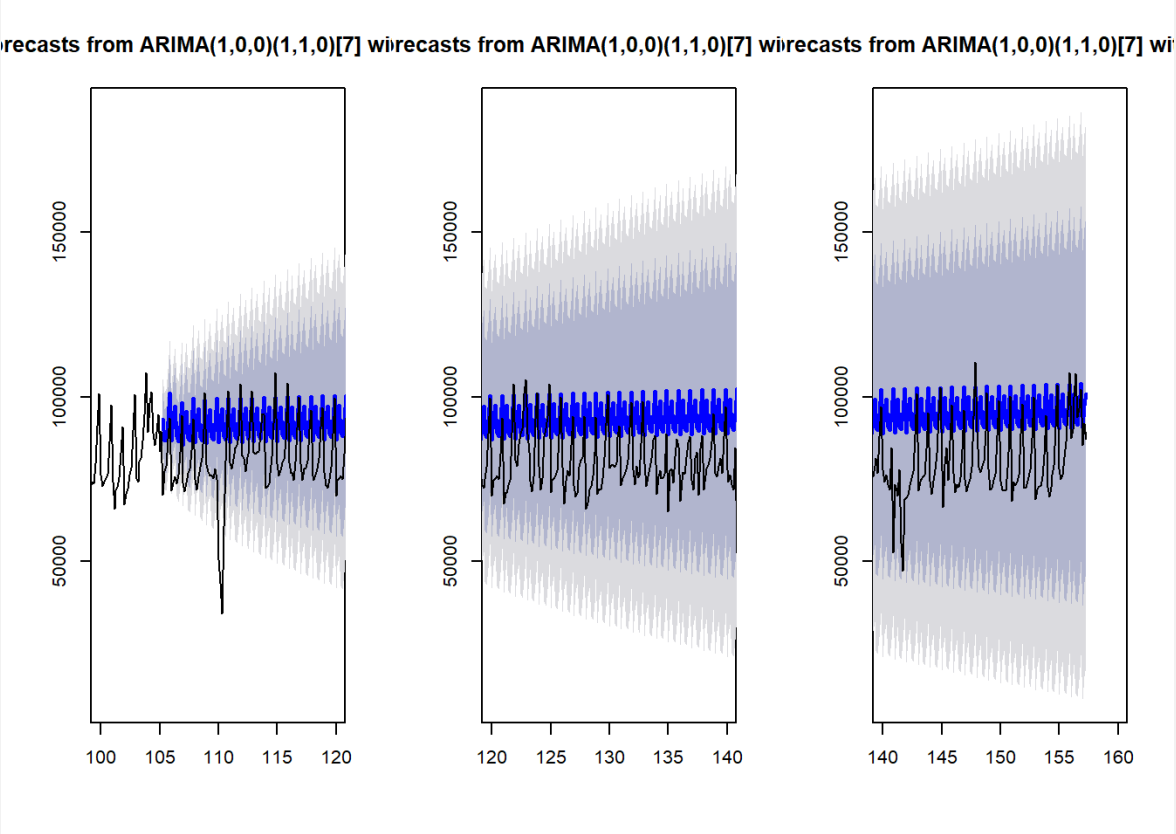


- 7일(1주)를 주기로 하였을 때,
  seasonal 성분과 아닌 성분의 분리가 잘 이루어진다.

- auto.arima로 ARIMA(1,0,1)(0,1,1)[7] 모델 적합

03 모델 선택 및 예측

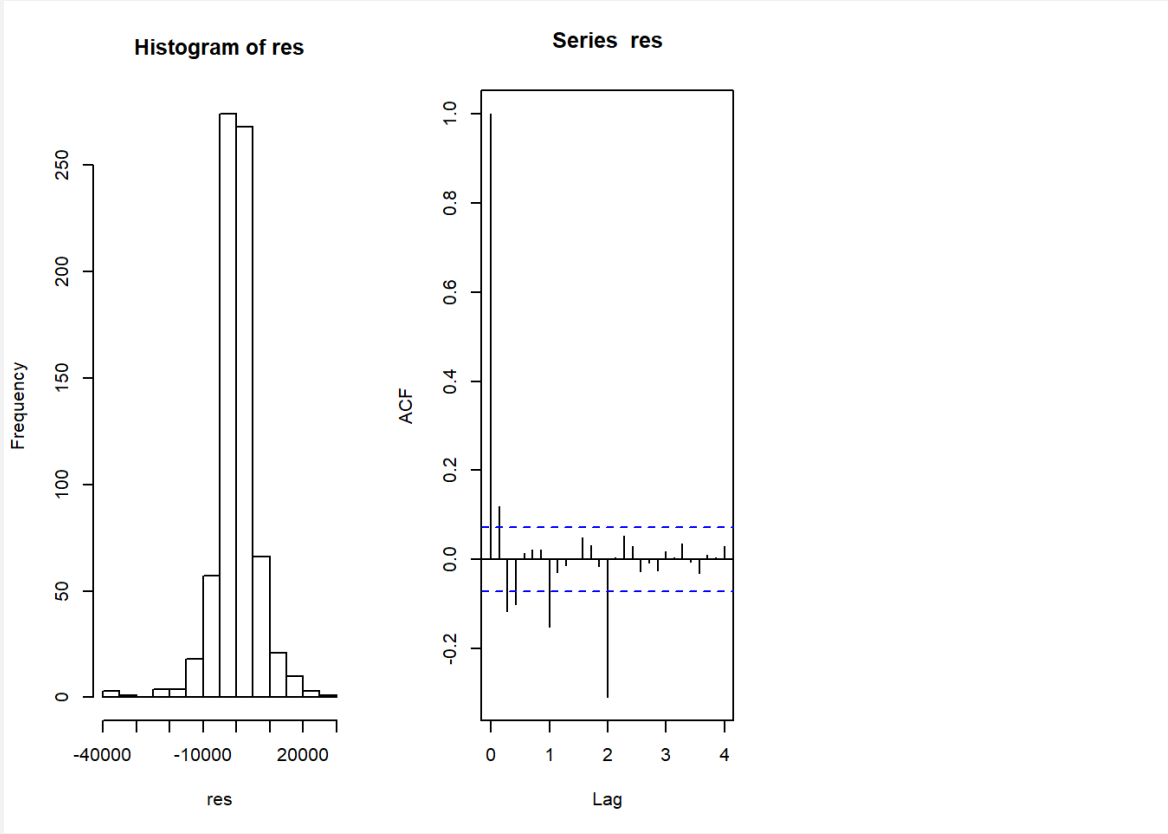
Group 2(홍대입구) : 승차인원이 평일보다 금,토에 많은 group



- 1년치를 예측하였을 때, 실제 데이터와 오차가 클 뿐더러 예측치의 분산이 점점 커져 의미가 없다.

03 모델 선택 및 예측

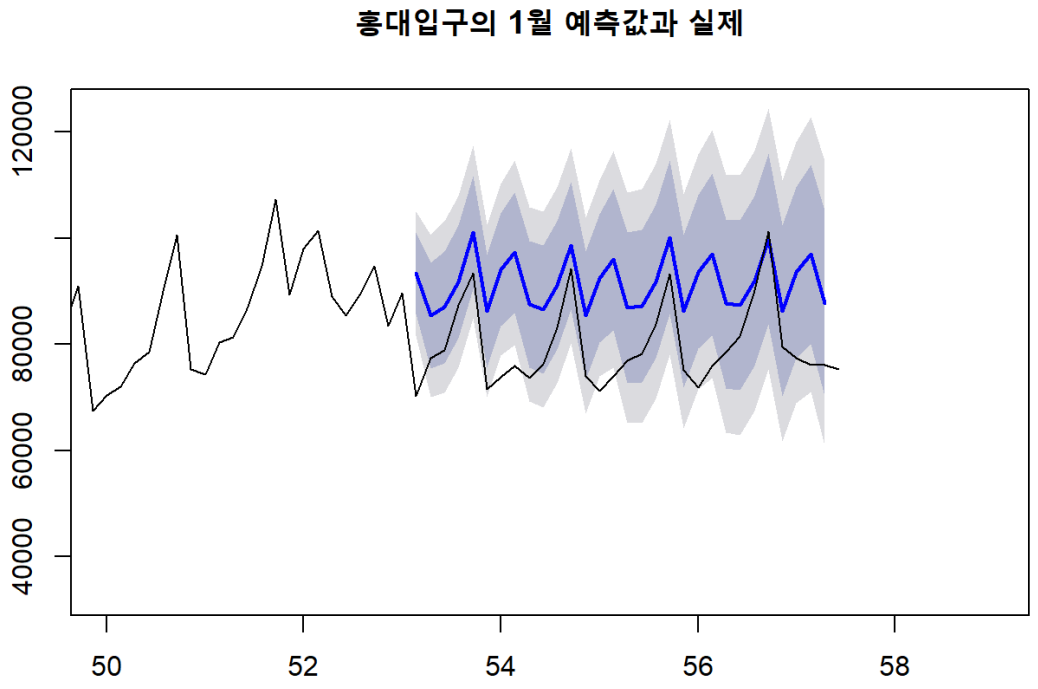
Group 2(홍대입구) : 승차인원이 평일보다 금,토에 많은 group

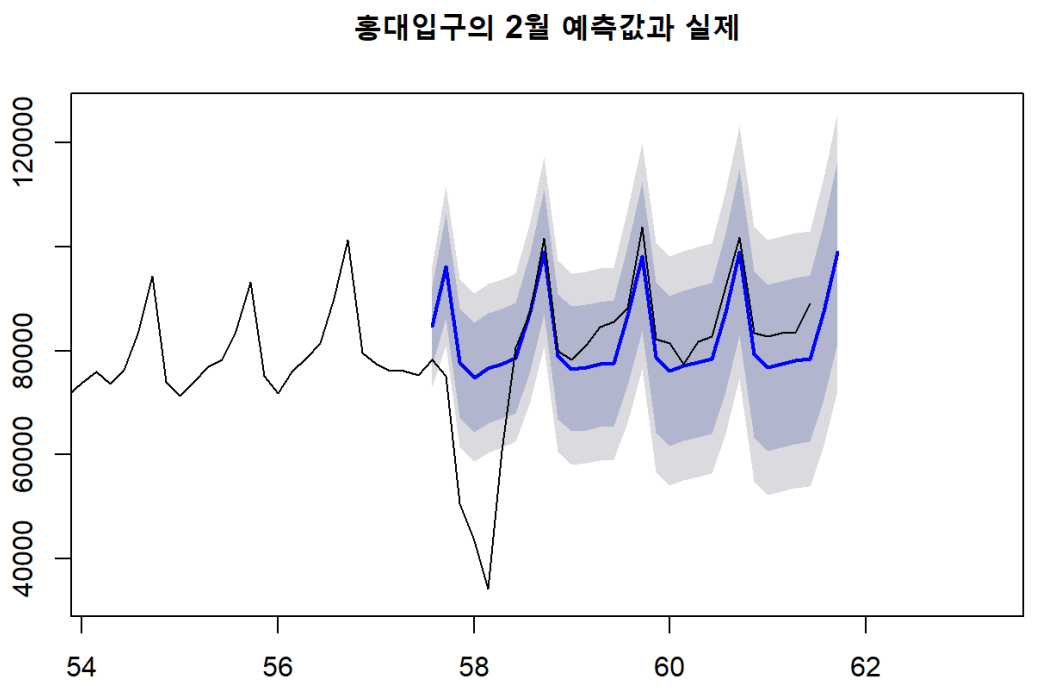


- 잔차의 자기상관성 또한 0이 아니다.
- 예측 범위를 줄여 월별 단위 예측을 시도하였다.
- 18년 데이터로 train했을 때 자기상관이 없다.
- ARIMA(2,0,2) (1,1,0) [7] 모델을 적합.

03 모델 선택 및 예측

Group 2(홍대입구) : 승차인원이 평일보다 금,토에 많은 group

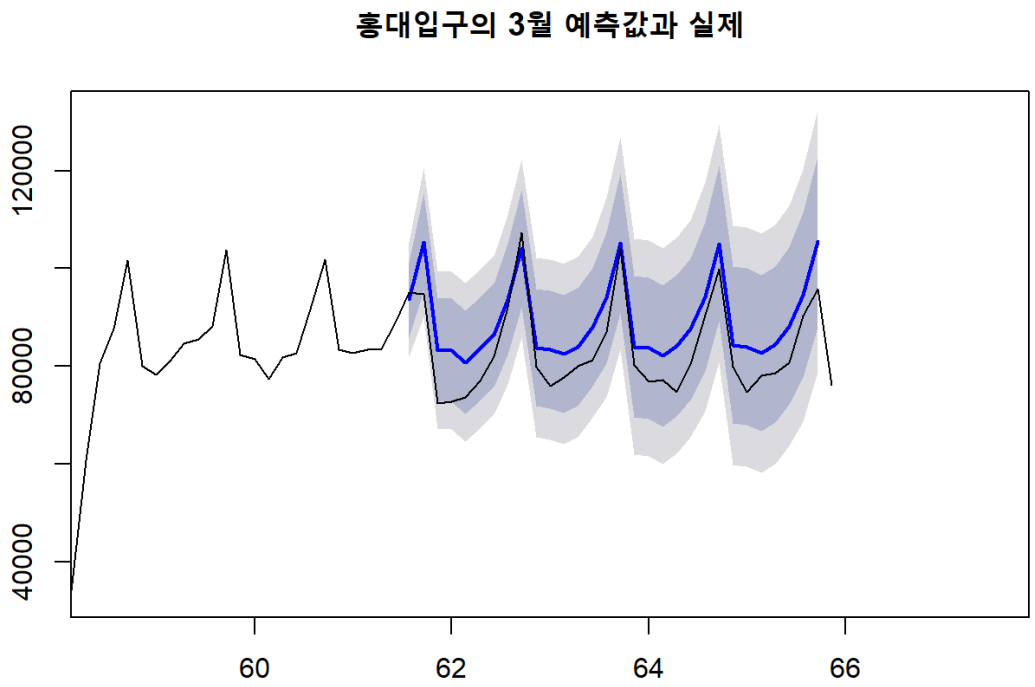


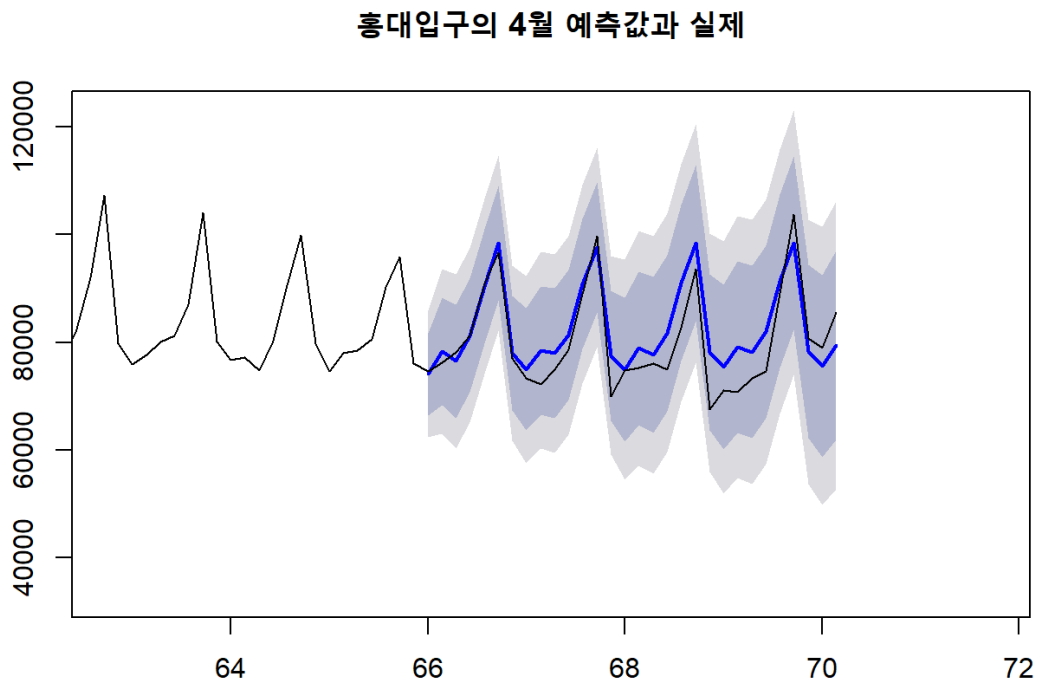


- 12월의 데이터에 영향을 받아 1월을 전반적으로 높게 예측하는 경향을 보인다.

03 모델 선택 및 예측

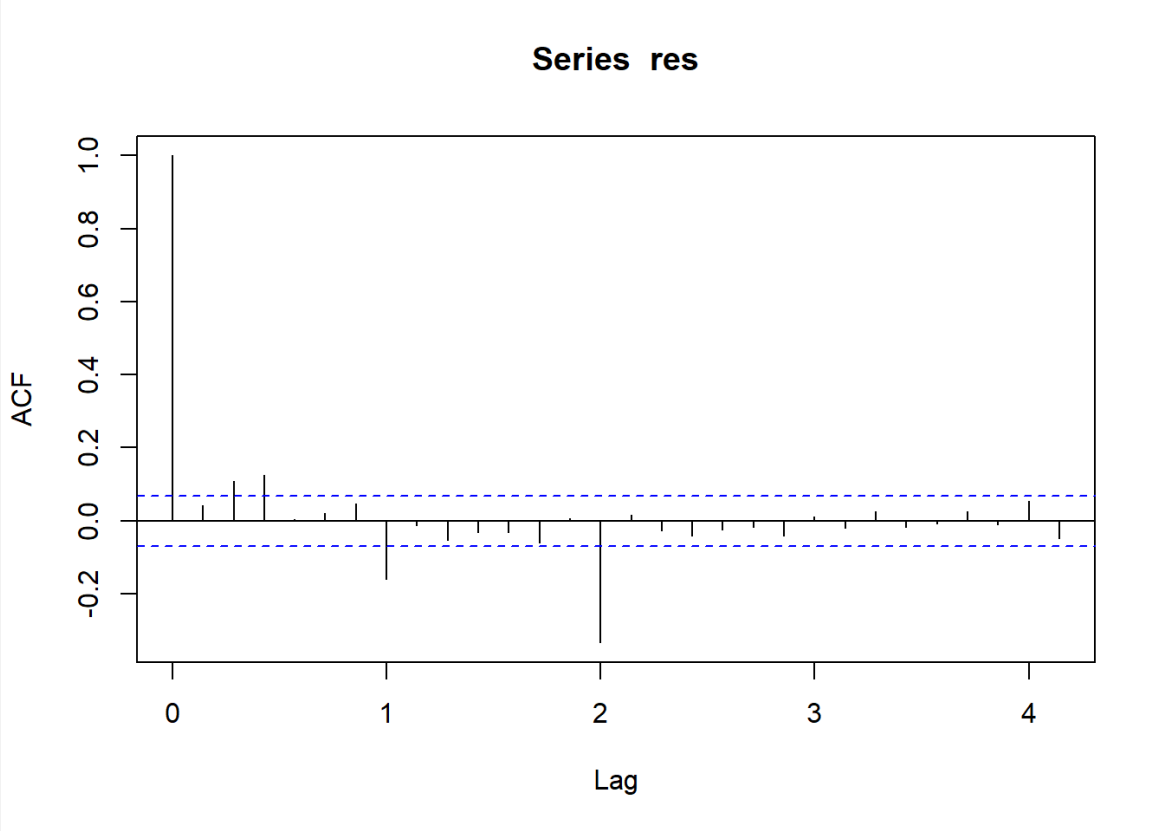
Group 2(홍대입구) : 승차인원이 평일보다 금,토에 많은 group





03 모델 선택 및 예측

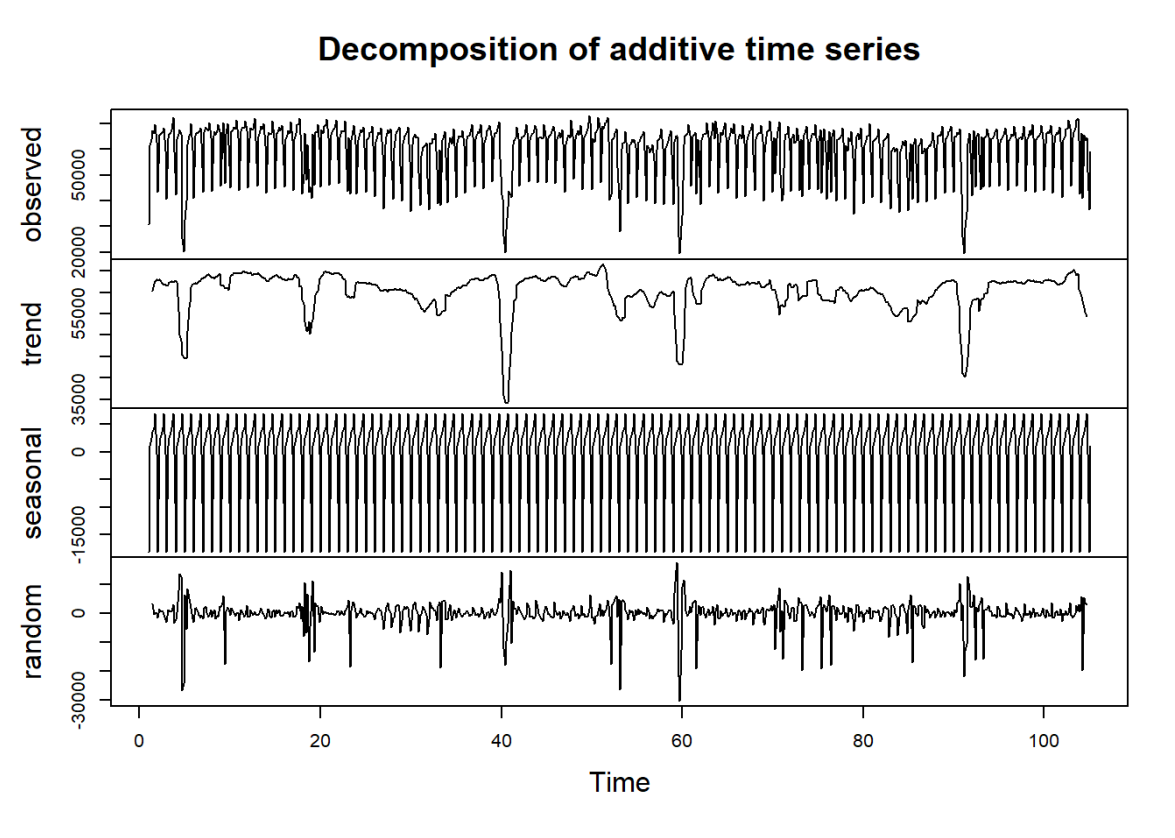
Group 2(홍대입구) : 승차인원이 평일보다 금,토에 많은 group



- 잔차의 acf 가 0에 가까운 값이 나온다.
- Box_Ljung test에서도 잔차의 자기상관성이 없다.
- Augmented DF test도 통과하여, stationary한 모델임을 알 수 있다.

03 모델 선택 및 예측

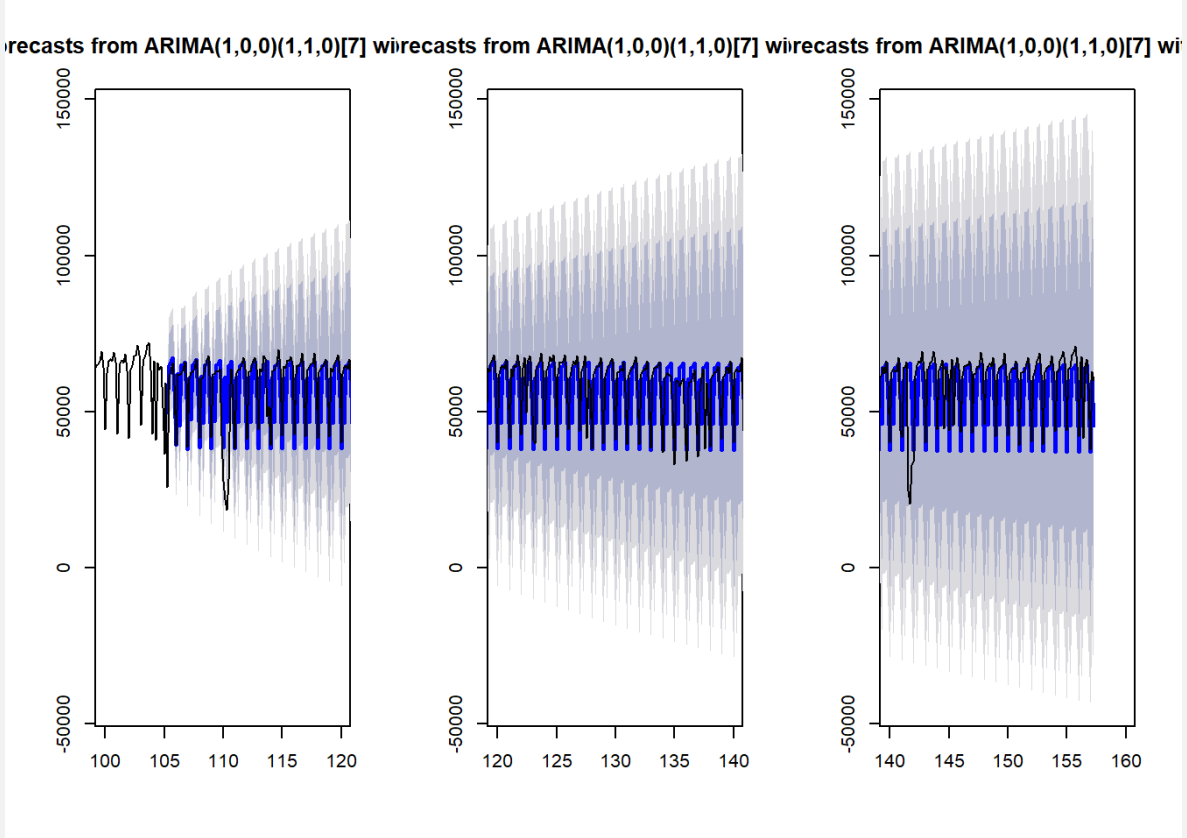
Group 3(신도림) : 일요일만 승차인원이 적은 group



- frequency를 7일(1주)로 하였다.
- auto.arima로 ARIMA(1,0,0)(1,1,0)[7] 모델 적합

03 모델 선택 및 예측

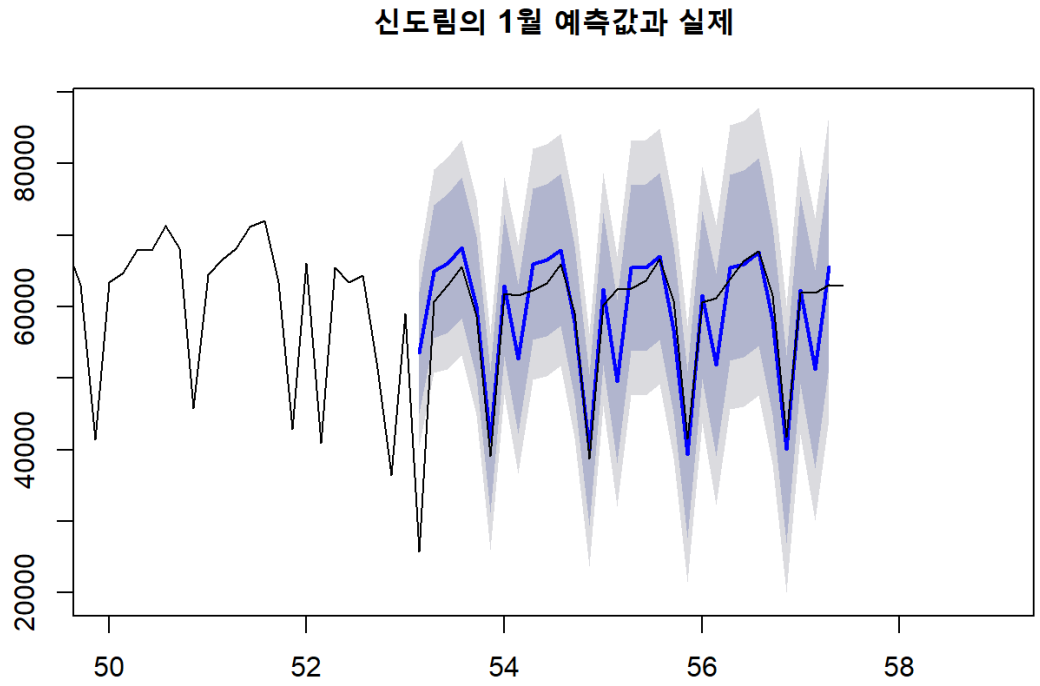
Group 3(신도림) : 일요일만 승차인원이 적은 group

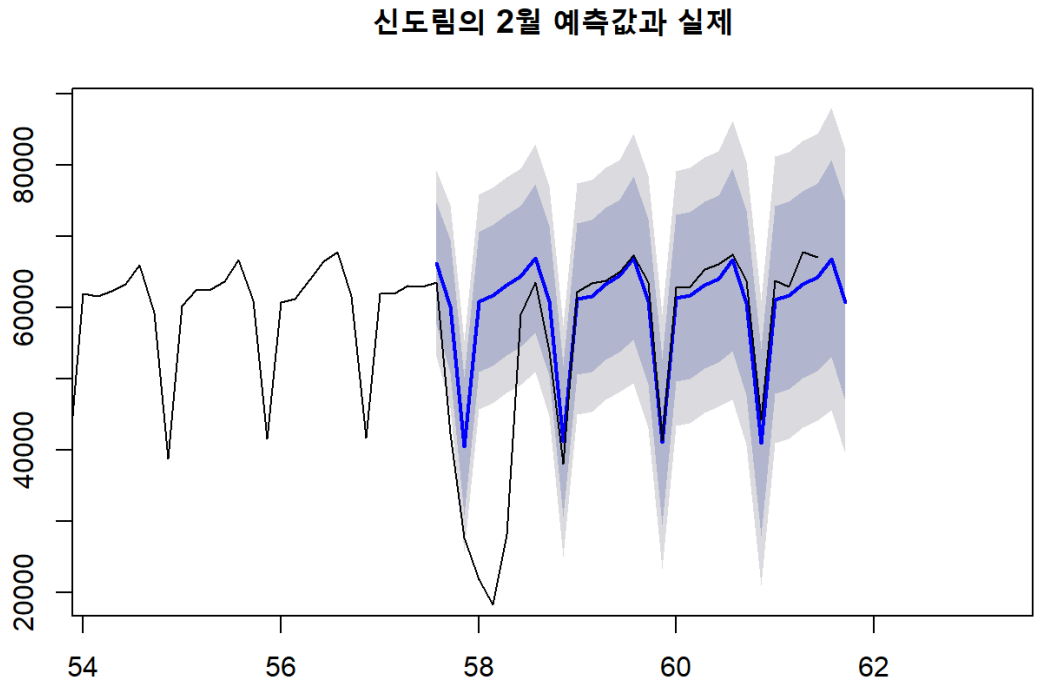


- 예측도 잘 맞지 않고 끝으로 갈 수록 분산이 커져 예측의 의미가 없다.
- 예측의 분산을 줄이기 위해 월별 예측을 진행하였다.
- 18년 데이터로 train했을 때 자기상관이 없다.
- ARIMA(2,0,0) (2,1,0) [7] 모델을 적합.

03 모델 선택 및 예측

Group 3(신도림) : 일요일만 승차인원이 적은 group

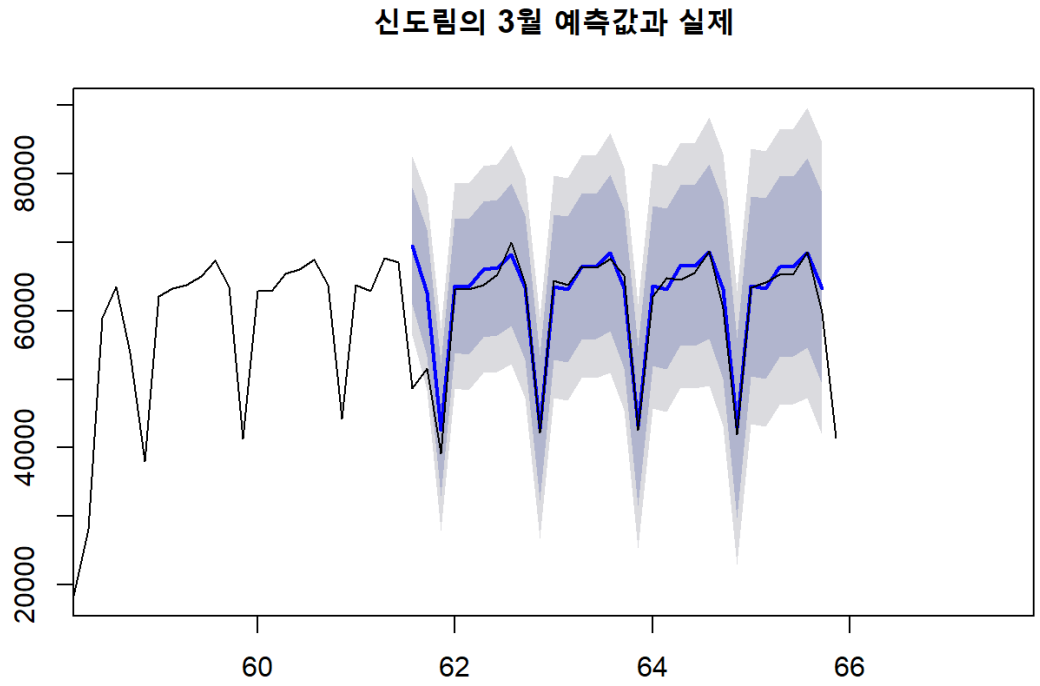


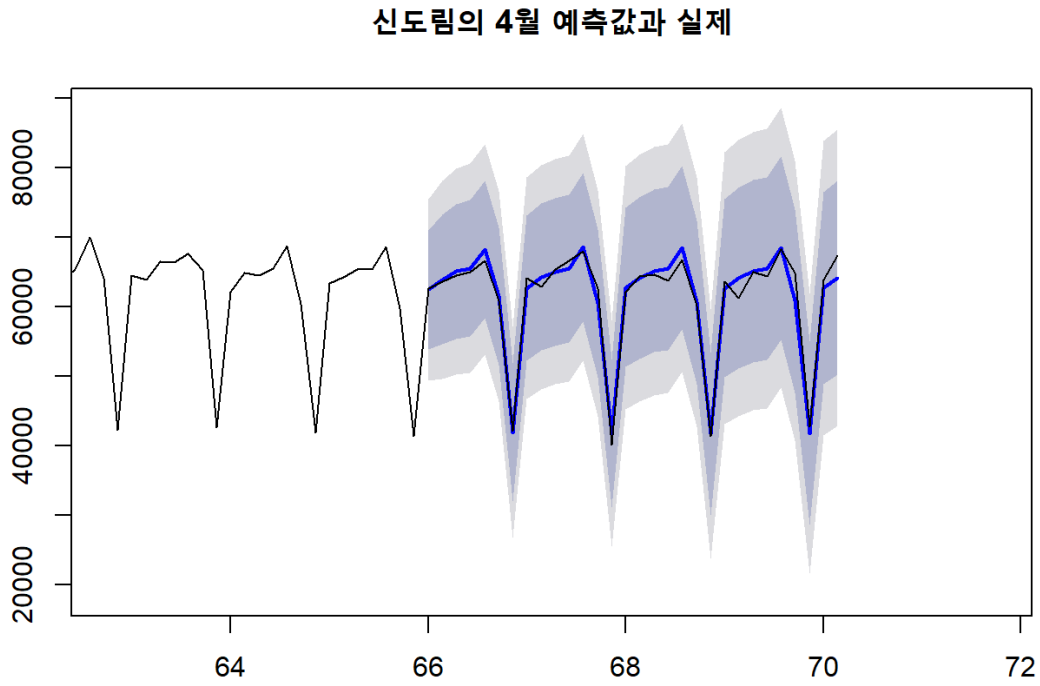


- 12월의 irregular한 데이터에 영향을 받아 1월은 불안정한 예측을 보인다.

03 모델 선택 및 예측

Group 3(신도림) : 일요일만 승차인원이 적은 group

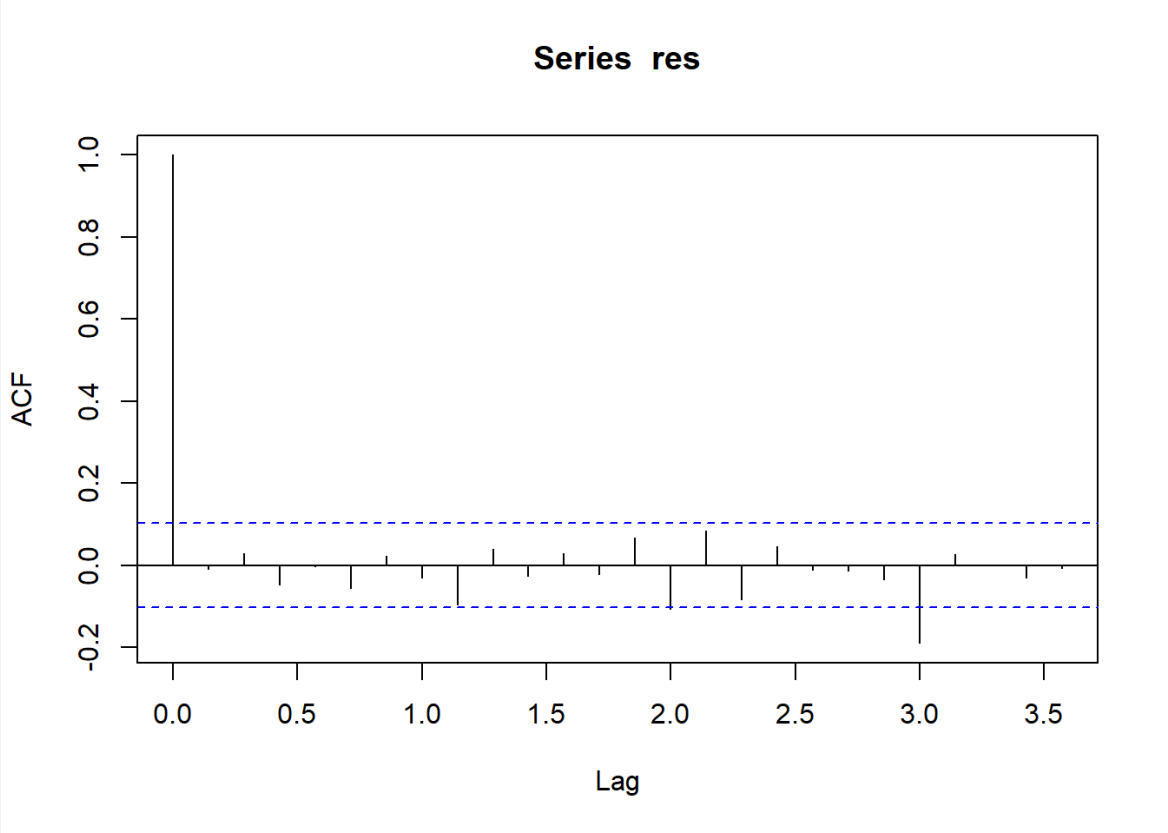




- 그다지 불규칙한 이벤트가 없는 3,4월의 예측이 더 좋다.

03 모델 선택 및 예측

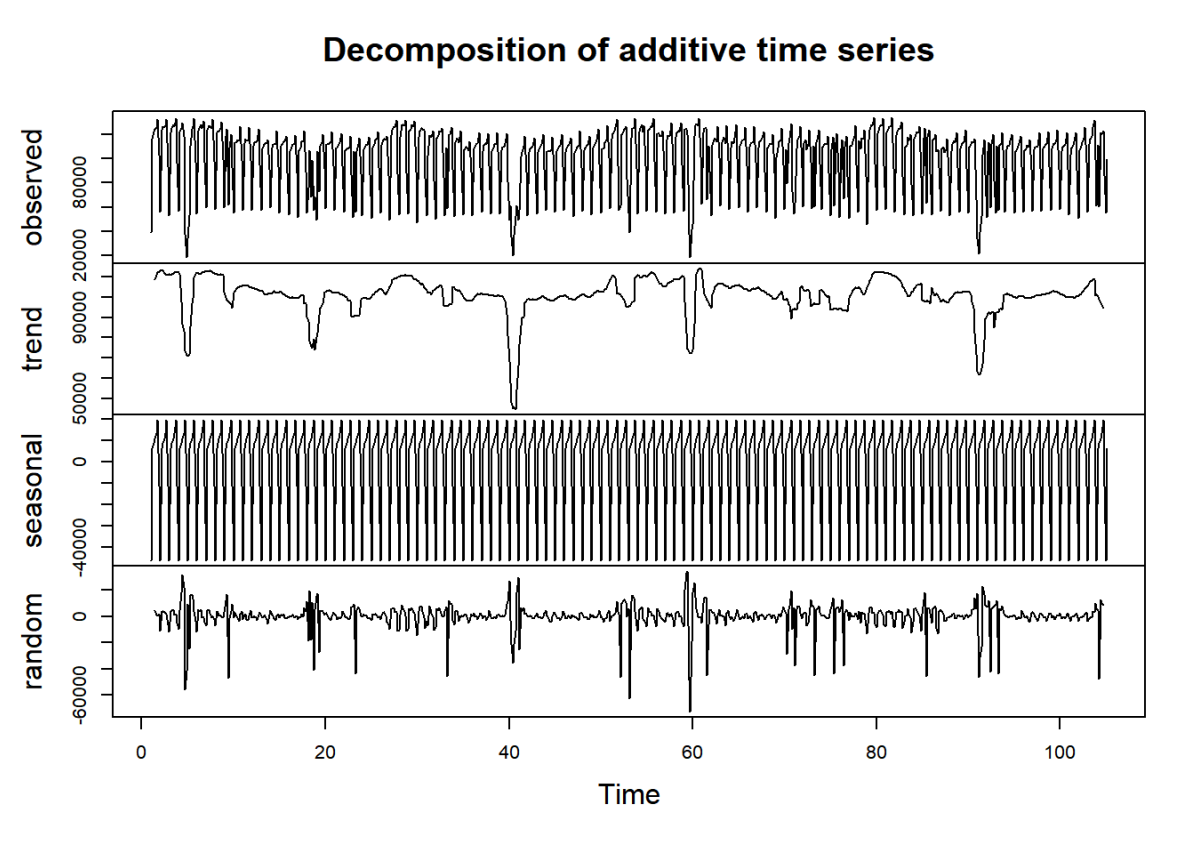
Group 3(신도림) : 일요일만 승차인원이 적은 group



- 잔차의 자기상관도 0 에 가깝다.
- Box-Ljung test에서도 p-value가 0.05 이상 나와 잔차의 자기상관이 없다고 할 수 있다.
- Augmented DF test도 통과하여, stationary한 모델임을 알 수 있다.

03 모델 선택 및 예측

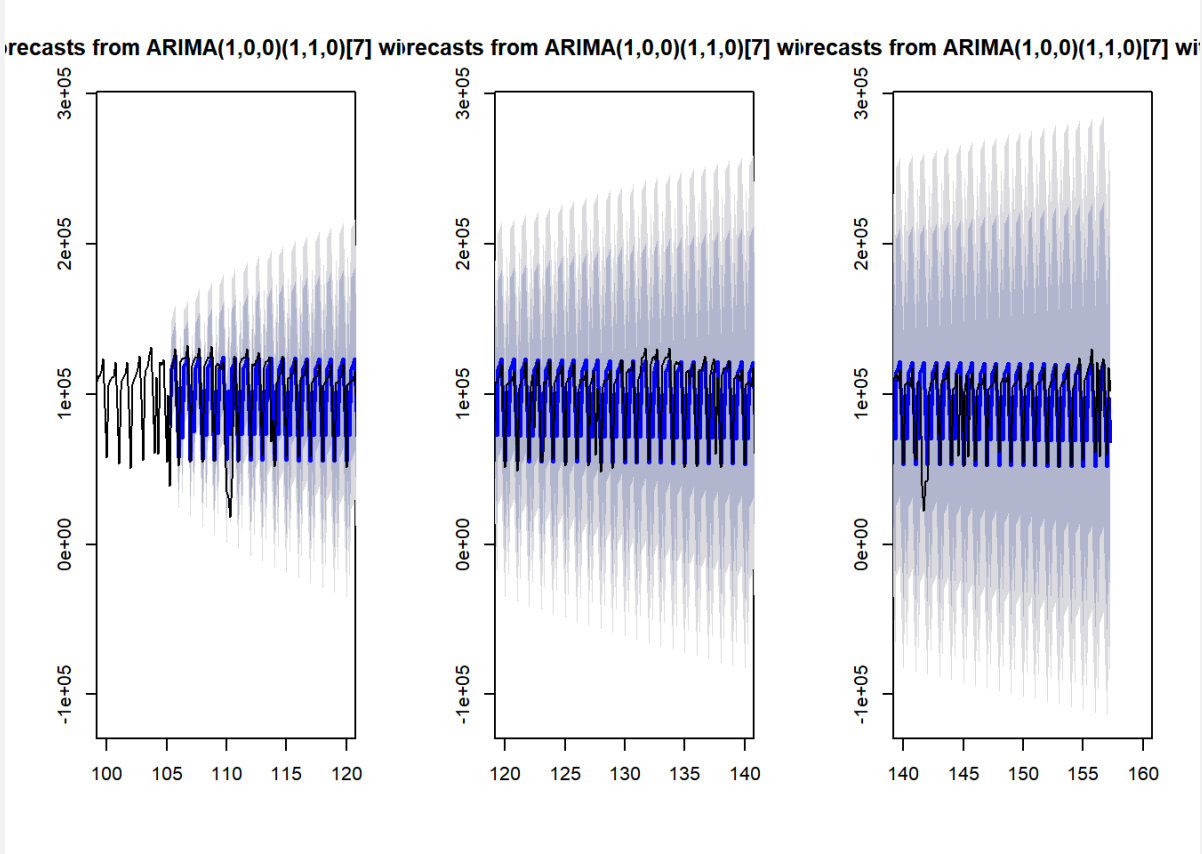
Group 4(강남) : 일요일의 승차인원이 가장 적고 토요일이 중간 정도인 group



- 주기 7을 갖는 seasonal 모델을 만들었다.

03 모델 선택 및 예측

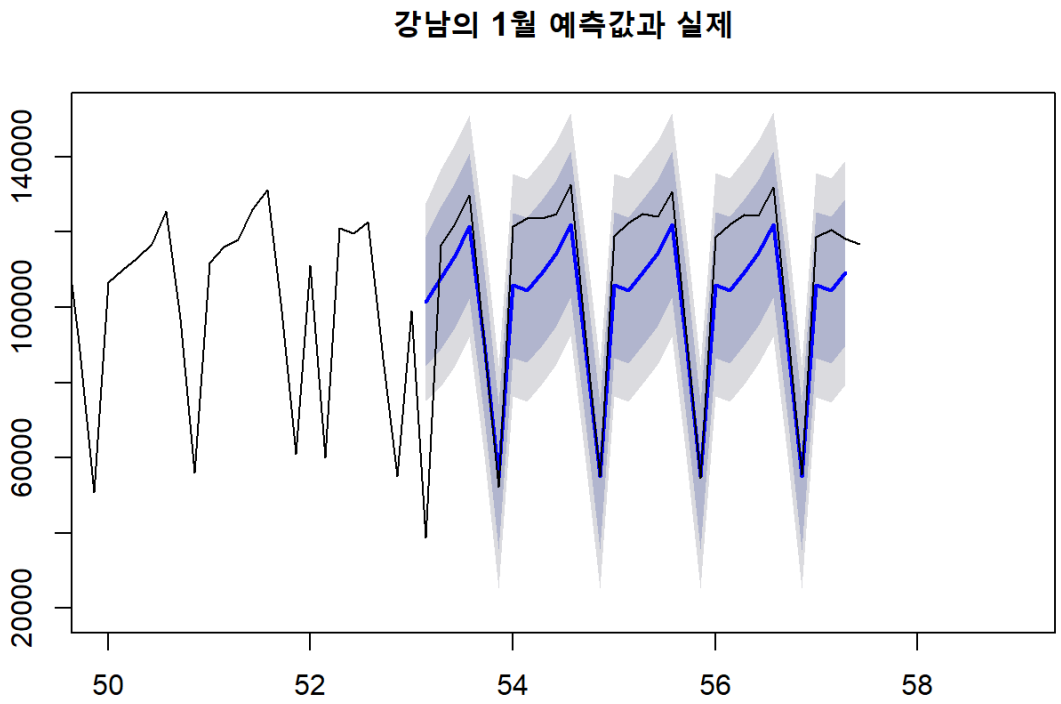
Group 4(강남) : 일요일의 승차인원이 가장 적고 토요일이 중간 정도인 group

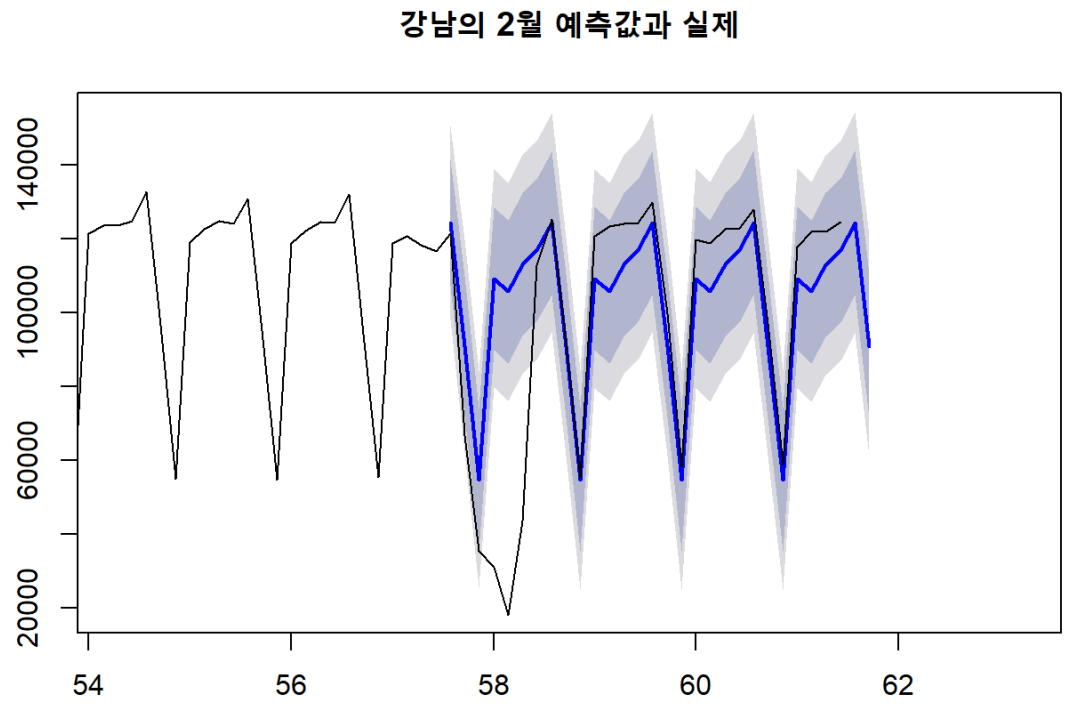


- ARIMA(1,0,1)(0,1,1)[7] 모델로 적합하였지만, 마찬가지로 1년치의 예측은 좋지 못하다.
- 예측의 분산을 줄이기 위해 월별 예측을 진행하였다.
- 18년 데이터로 train했을 때 자기상관이 없다.
- ARIMA(2,0,1)(0,1,1)[7] 모델을 적합.

03 모델 선택 및 예측

Group 4(강남) : 일요일의 승차인원이 가장 적고 토요일이 중간 정도인 group

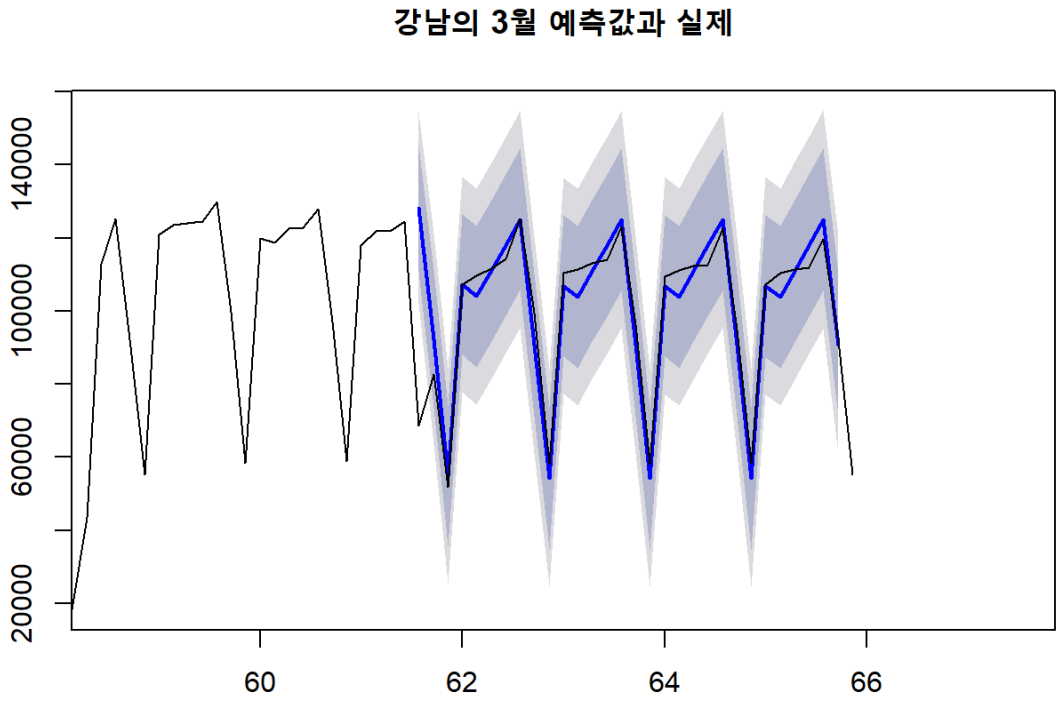


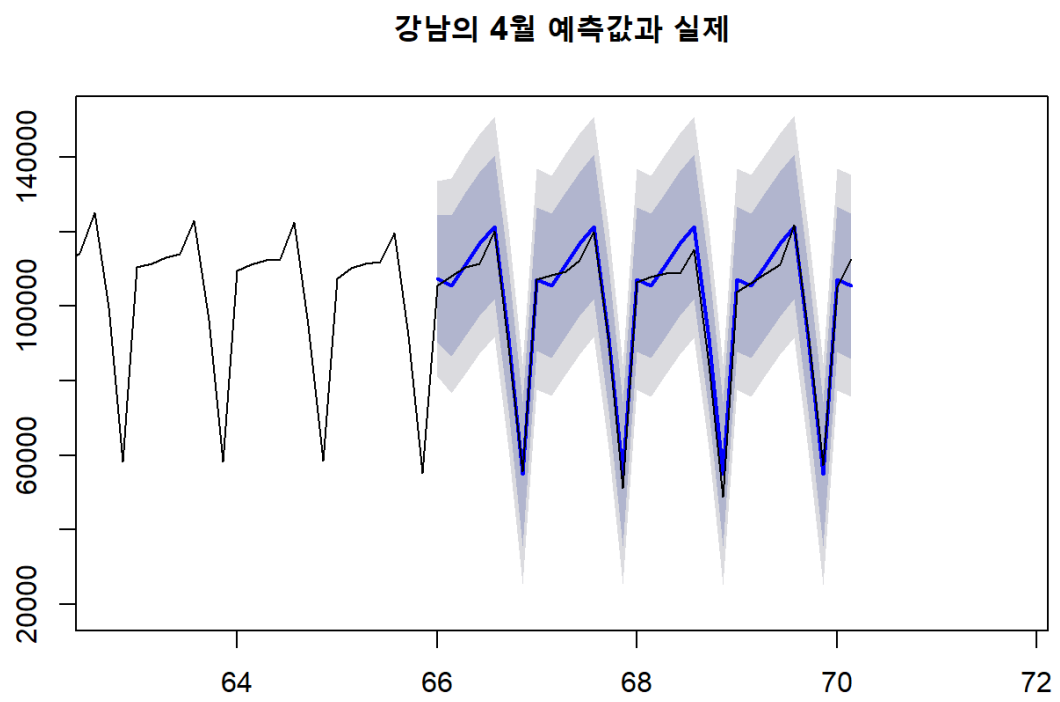


- 12월의 irregular한 데이터에 영향을 받아 1월의 값을 잘 예측하지 못한다. 1월의 승객 수가 평소보다 많다.

03 모델 선택 및 예측

Group 4(강남) : 일요일의 승차인원이 가장 적고 토요일이 중간 정도인 group

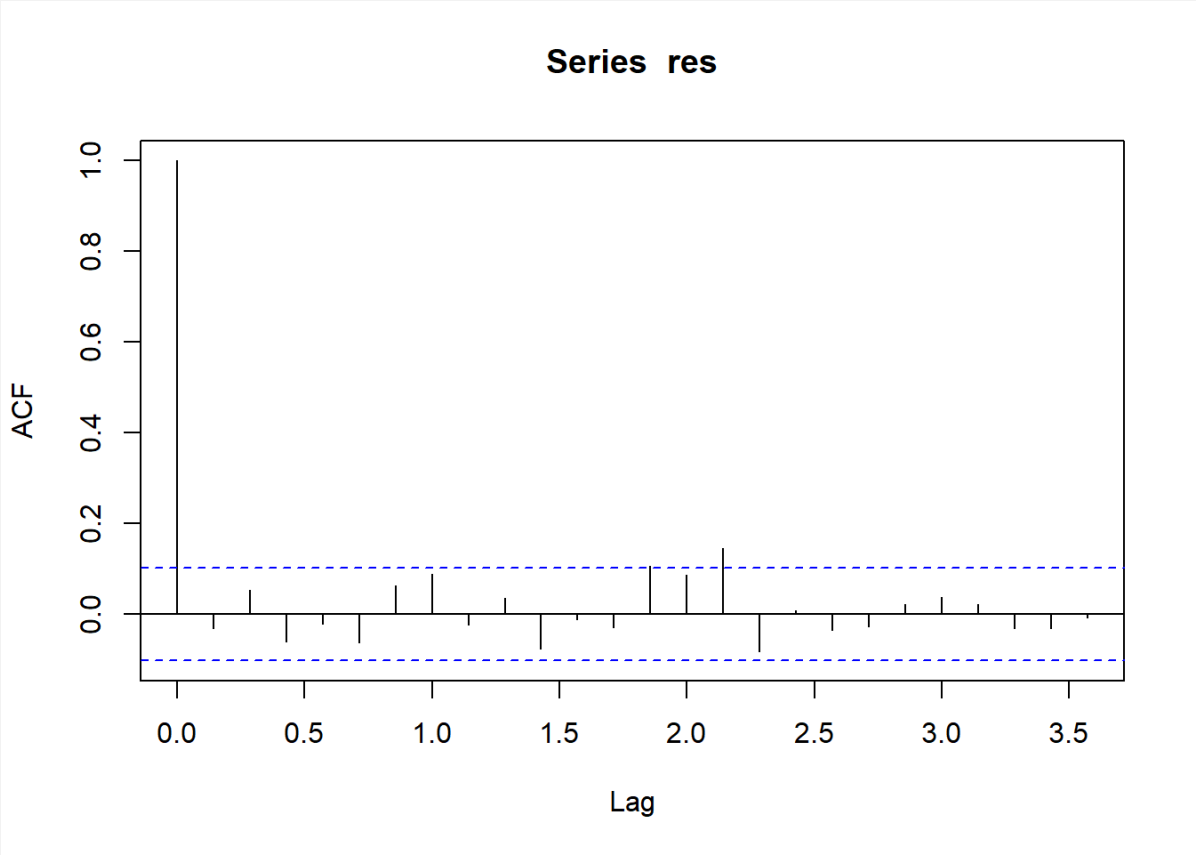




- 그다지 불규칙한 이벤트가 없는 3,4월의 예측이 더 좋다.

03 모델 선택 및 예측

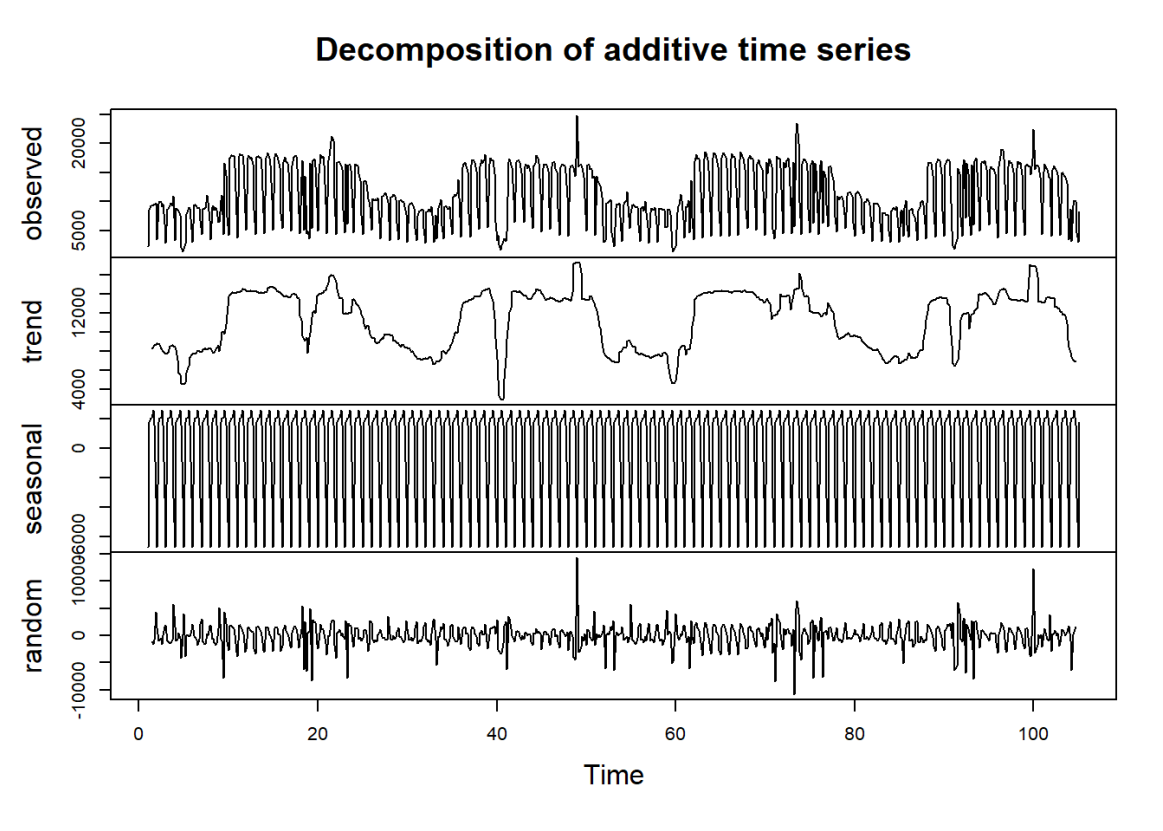
Group 4(강남) : 일요일의 승차인원이 가장 적고 토요일이 중간 정도인 group



- 잔차의 자기상관도 0 에 가깝다.
- Box-Ljung test에서도 p-value가 0.05 이상 나와 잔차의 자기상관이 없다고 할 수 있다.
- Augmented DF test도 통과하여, stationary한 모델임을 알 수 있다.

03 모델 선택 및 예측

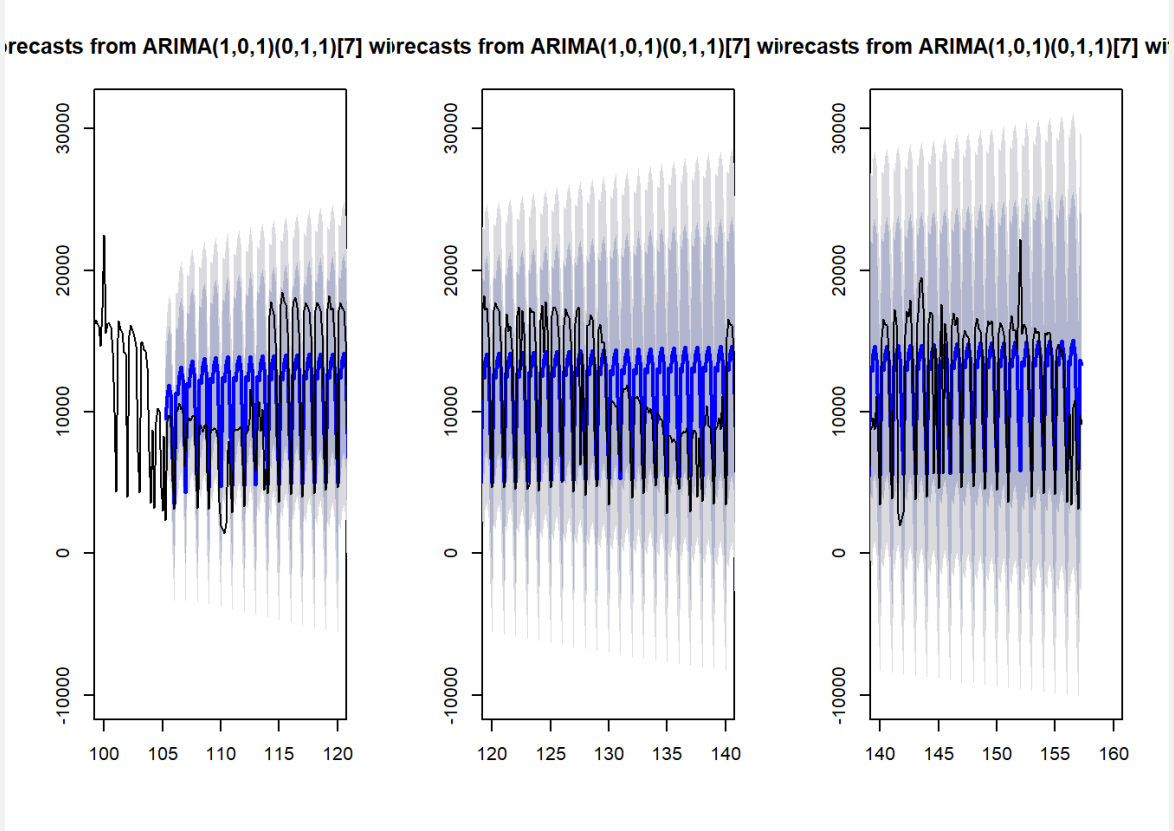
Group 5(한양대입구) : 특정 기간에 승차인원이 많은 group



- findfrequency를 함수를 썼을 때 1주의 주기는 찾아내 주지만, 6개월의 주기(한 학기와 방학)는 찾아내주지 못한다.
- 6개월의 주기와 1주의 주기를 한꺼번에 적합하기에는 데이터가 부족.
- auto.arima로 ARIMA(1,0,1)(0,1,1)[7] 모델 적합

03 모델 선택 및 예측

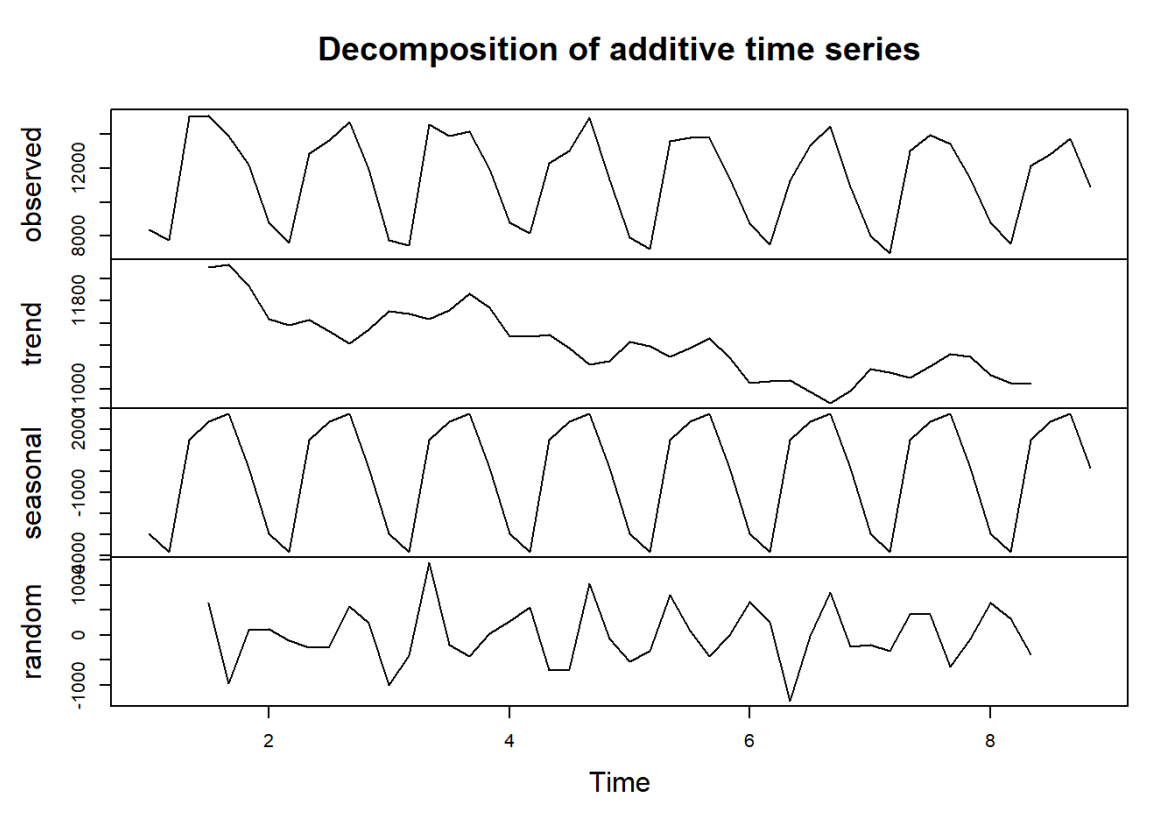
Group 5(한양대입구) : 특정 기간에 승차인원이 많은 group



- 특정 기간에 따른 승차인원의 많고 적음을 제대로 예측하지 못한다.

03 모델 선택 및 예측

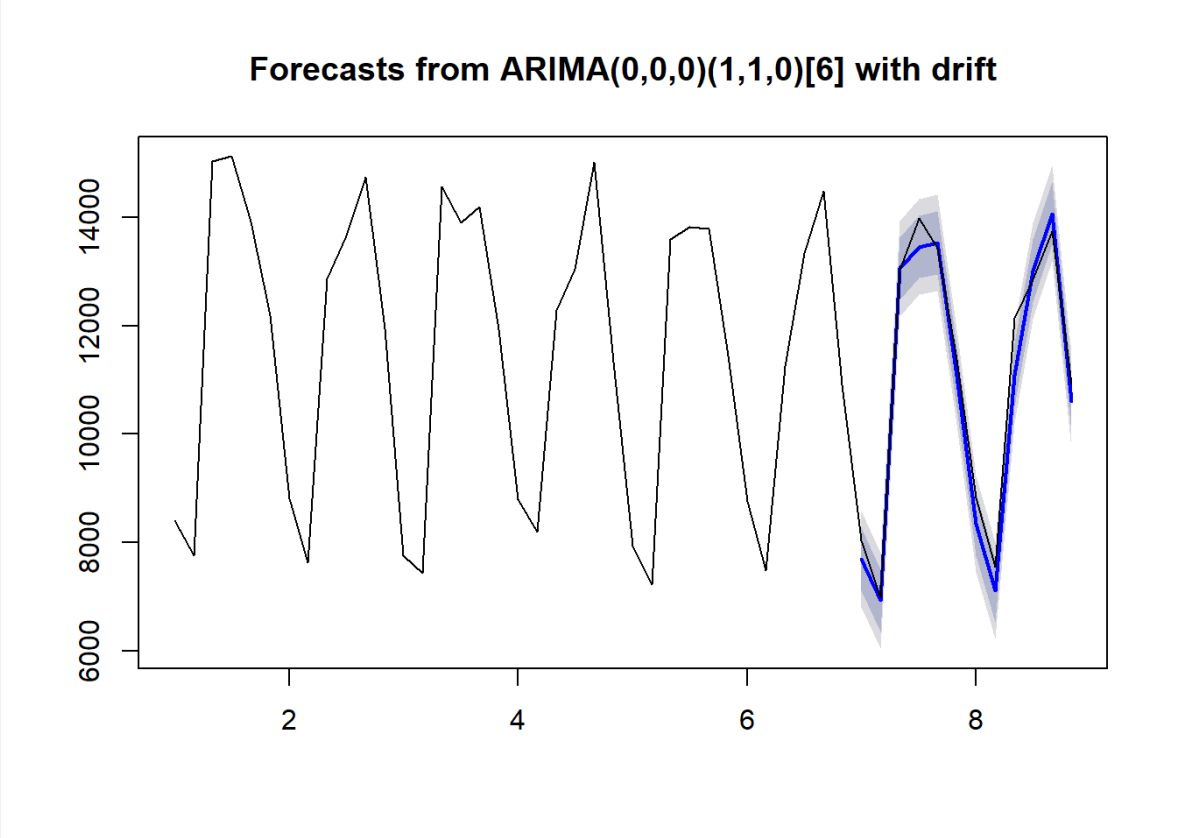
Group 5(한양대입구) : 특정 기간에 승차인원이 많은 group



- 월별 평균 승차인원 수로, 월 단위 예측 시도
- findfrequency 함수에서 주기가 6이 나왔다(학기+방학).
- auto.arima로 ARIMA(0,0,0)(1,1,0)[6] 모델 적합
- 18년까지의 월별 데이터로, 19년의 월별 평균 승차인원 예측

03 모델 선택 및 예측

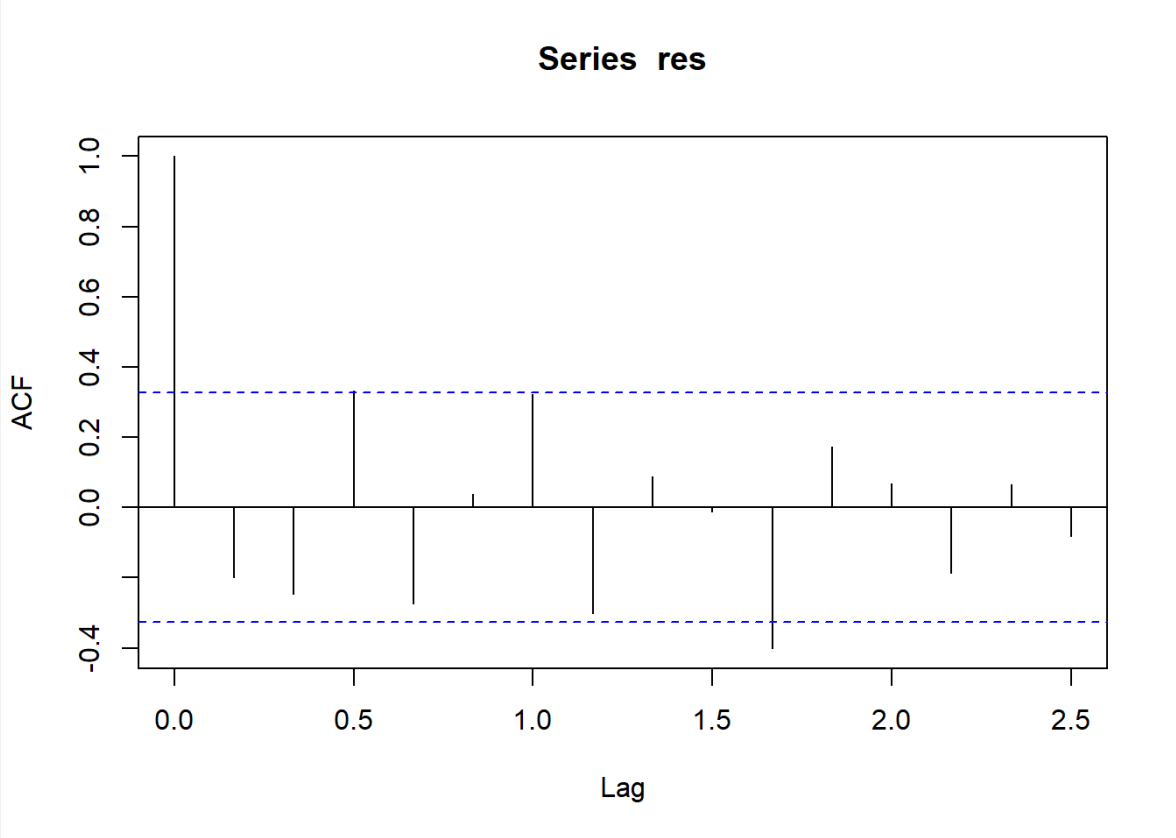
Group 5(한양대입구) : 특정 기간에 승차인원이 많은 group



- 예측이 꽤 잘 이루어짐을 볼 수 있다.

03 모델 선택 및 예측

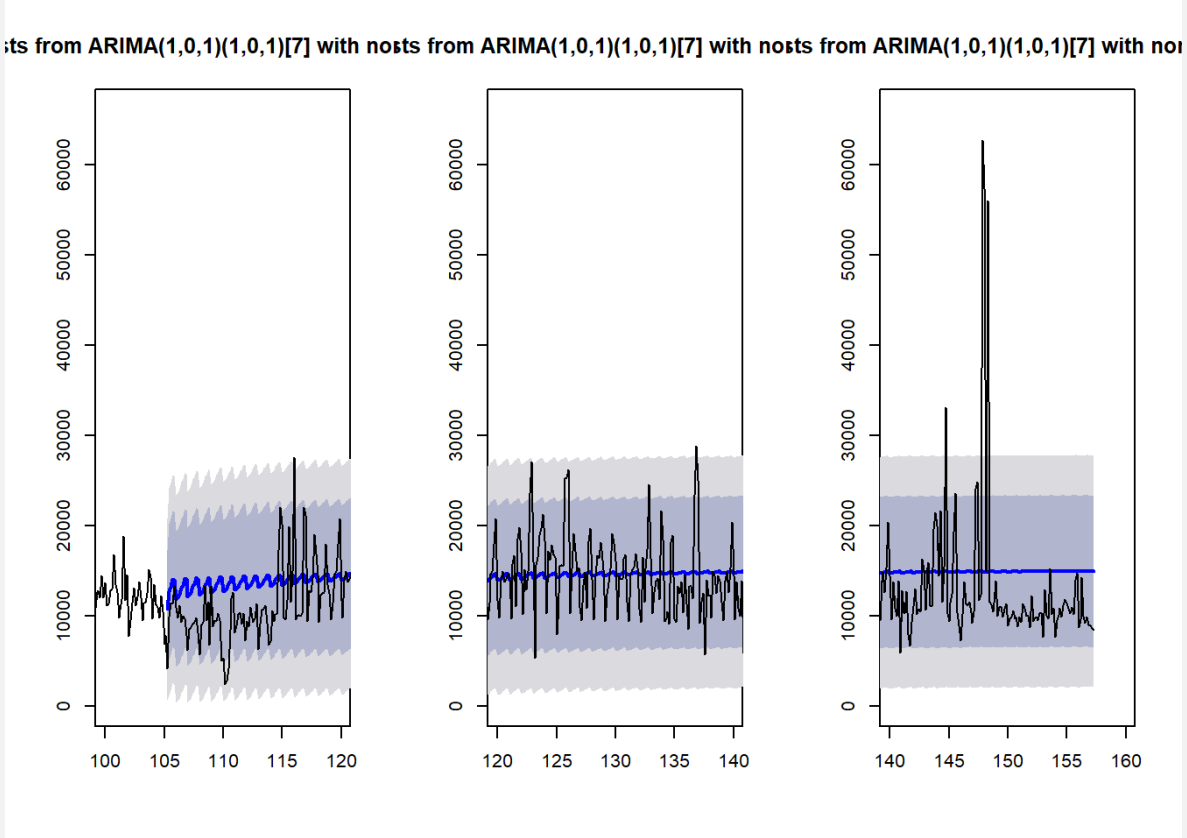
Group 5(한양대입구) : 특정 기간에 승차인원이 많은 group



- 잔차의 acf 가 0에 가까운 값이 나온다.
- Box_Ljung test에서도 잔차의 자기상관성이 없다.

03 모델 선택 및 예측

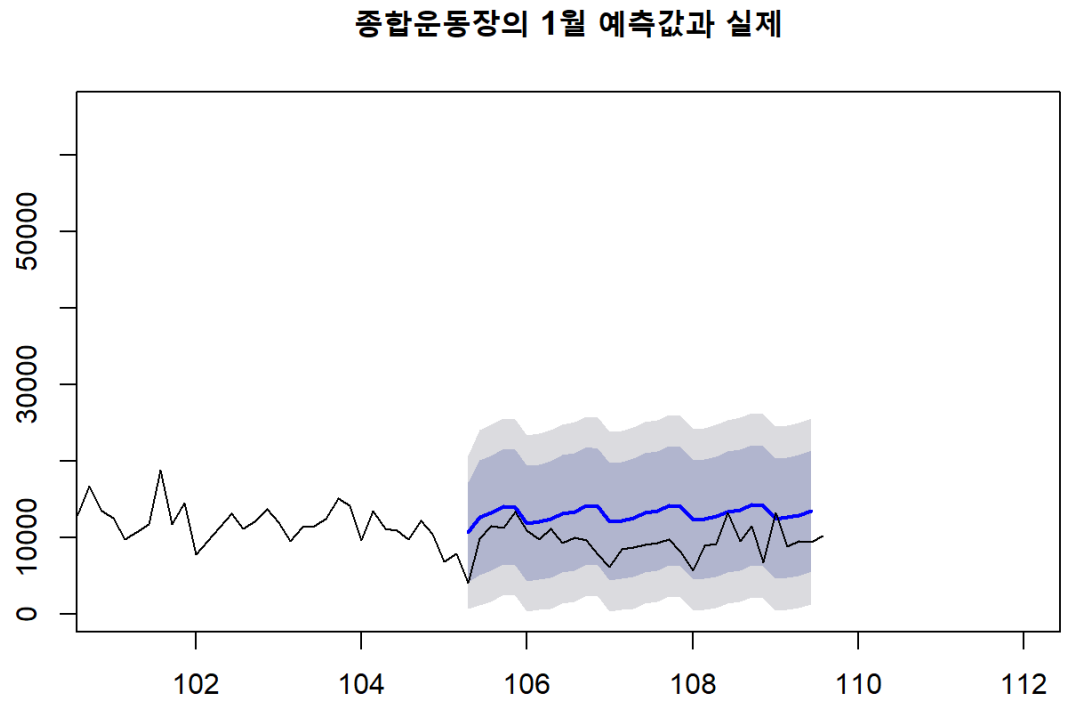
Group 6(종합운동장) : 특별한 날에 승차인원이 많은 group

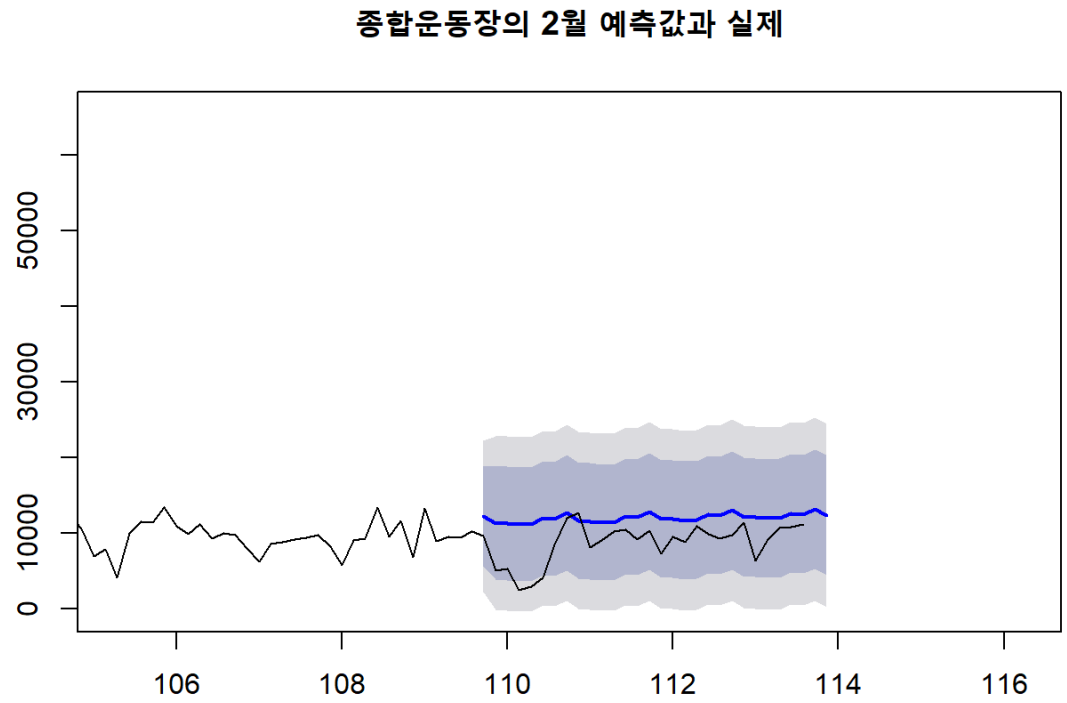


- ARIMA(1,0,1)(0,1,1)[7]모델을 적합시켜 보았지만 예측의 정확도가 크게 떨어지는 것을 볼 수 있다.

03 모델 선택 및 예측

Group 6(종합운동장) : 특별한 날에 승차인원이 많은 group

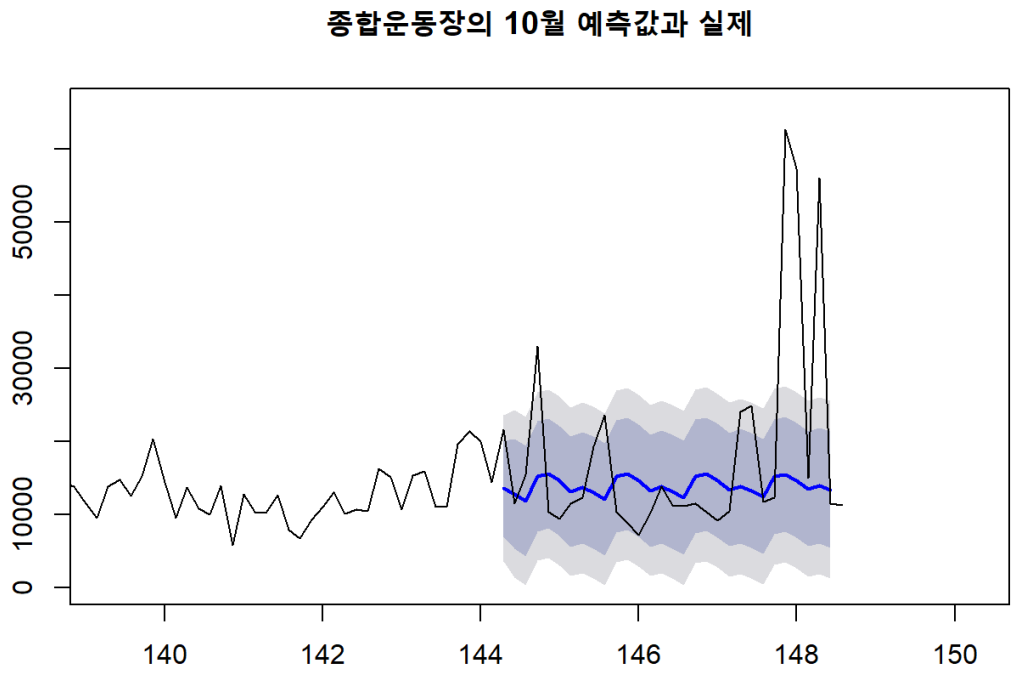


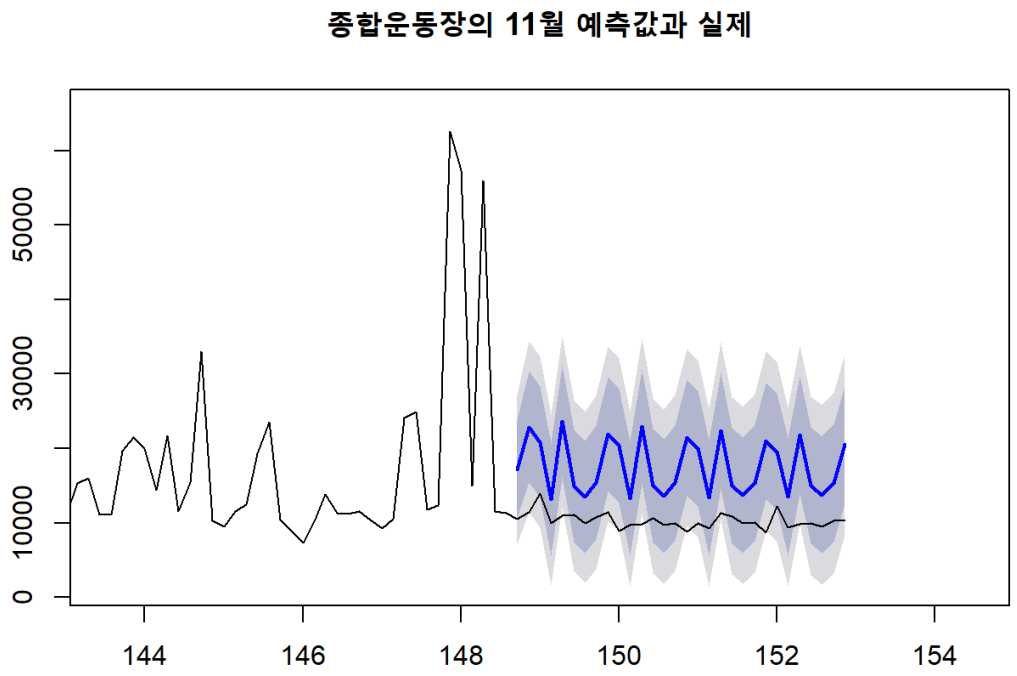


- 한 달 단위로 끊어 모델을 만들어 예측을 하였을 때도, 그다지 좋은 예측을 하지 못 한다.

03 모델 선택 및 예측

Group 6(종합운동장) : 특별한 날에 승차인원이 많은 group





- 특히, 데이터에 이상치가 있었을 때 영향을 굉장히 많이 받아 robust하지 못한 결과물을 보여준다.

04 결론

	Group1 시청	Group5 한양대입구(월별)	Group6 종합운동장
17,18년 데이터로 적합	ARIMA(1,0,1)(0,1,1)[7]	ARIMA(0,0,0)(1,1,0)[6]	ARIMA(2,0,1)(0,1,1)[7]

	Group2 홍대입구	Group3 신도림	Group4 강남
18년 데이터로 적합	ARIMA(2,0,2)(1,1,0)[7]	ARIMA(1,0,0)(1,1,0)[7]	ARIMA(2,0,1)(0,1,1)[7]

- 17,18년 데이터로 적합해보고, 모델의 자기상관성이 있으면 18년 데이터로만 적합을 하였다.

04 결론

- Group1. 시청
- 주기성이 강하고 outlier들이 많지 않아 예측이 잘 된다.
- 주로 규칙적으로 출근하는 사람들이 많이 이용하는 역이라 그럴 것이라 생각된다.

- Group2. 홍대입구
- 불규칙적인 데이터가 많아 예측이 어렵다.
- 대표적인 유흥지인 만큼 날에 따라 이용하는 승객의 편차가 크기 때문일 것이라 생각된다.

- Group3. 신도림
- 특별한 이벤트가 없는 3,4월의 경우 예측이 상당히 좋다.
- 마찬가지로 규칙적으로 출근하는 사람들이 많이 이용하는 역이라 그럴 것이라 생각된다.

04 결론

- Group4. 강남
- 1,2월은 예측이 잘 안 되지만, 3,4월은 예측이 좋다
- 유흥가이면서 동시에 출근하는 직장인들이 많이 이용하는 역이라 그럴 것이라 생각된다.

- Group5. 한양대입구
- 월별 예측이 잘 된다.
- 학기에 따라 규칙적으로 움직이는 대학생들이 많이 이용하는 역이라 그럴 것이라 생각된다.

- Group6. 종합운동장
- 예측을 거의 하지 못한다.
- 불규칙한 행사가 많고, 극단적인 outlier가 포함되어 있어서 그럴 것이라 생각된다.