**데이터과학 텀프로젝트 기획안**



과목 : 데이터과학(8772)

담당 교수 : 신문선 교수님

학과 : 소프트웨어전공

학번 : 202020899

이름 : 박보겸

1. 주제

지하철 혼잡도를 분석하여 가장 혼잡한 시간대를 알아내 지하철 시간대 증가 등 대책을 마련할 수 있도록 한다.

1. 주제 선정 이유

최근 이태원 참사가 일어난 것을 보고 사람이 밀집된 곳이 얼마나 위험한지 심각성을 깨닫게 되었다. 사람 밀집으로 인한 압사로 사망자가 150명이 넘는 사건이였는데, 사실 할로윈 축제가 아니더라도 좁은 곳에서는 인구가 밀집되면 압사 사고가 일어날 수 있다. 따라서 이와 같은 참사가 다시는 발생하지 않도록 예방하고 대책을 마련해야한다. 할로윈 축제와 같은 특별한 행사가 아닌 평소에도 사고는 충분히 발생할 수 있다. 평소에 사람이 가장 몰리는 곳은 출,퇴근길의 지하철이라고 생각이 되어 지하철 혼잡도를 분석하여 대책을 마련할 수 있도록 하고싶었다.

1. 분석 방법

가장 인구가 많이 밀집되는(혼잡한) 시간대를 한 눈에 볼 수 있어야 하기에 데이터 시각화를 통해 분석하기로 하였다.

ggplot함수를 이용하여 시간대 별 혼잡도를 비교할 수 있도록 할 예정이며, 함수를 이용한 다양한 모양의 그래프로 나타내어 가장 비교가 쉬운 그래프는 어떤 것인지도 한 눈에 비교하여 볼 수 있도록 할 예정이다. 또한 역 별로 어느 곳에서 가장 사람들이 많이 몰리는지 출퇴근 시간별 혼잡도를 비교함으로써 해당 역에는 특정 지하철 칸에 사람이 몰려 더 혼잡해지지 않도록 분산되어 줄을 서라고 하는 등 도움이 될 수 있는 방안을 마련하는 데에 도움을 준다.

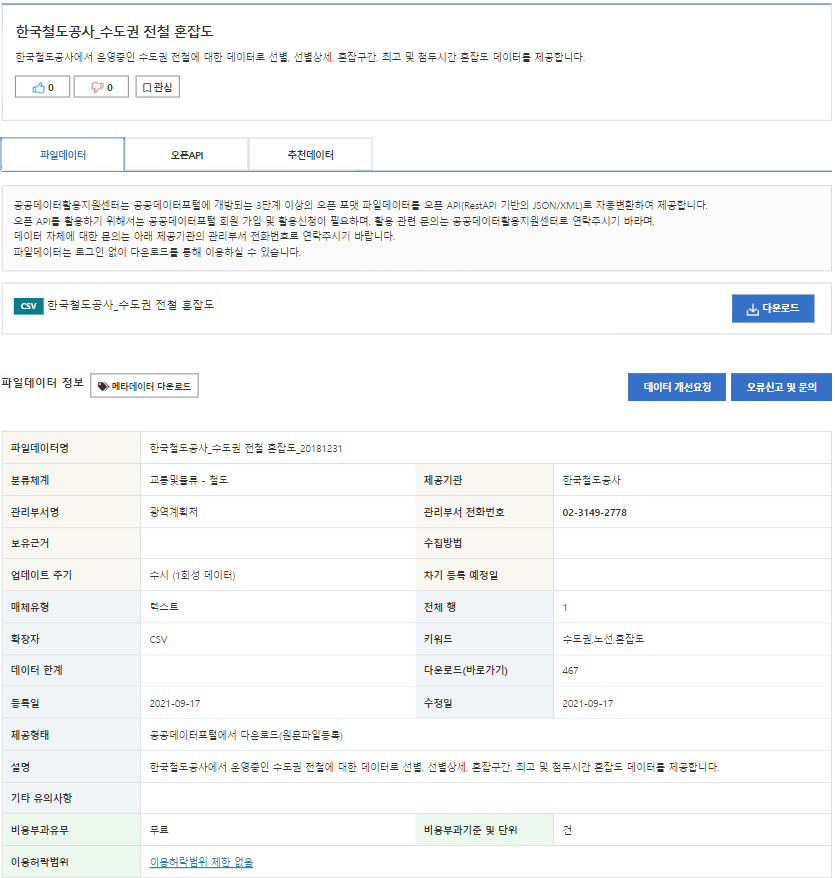
1. 데이터 수집

data <- read.csv(file="C:/Users/User/Documents/한국철도공사\_수도권 전철 혼잡도\_20181231.csv", header=T, fileEncoding = "euc-kr")

data

congestion <- data %>% select(c(최대.혼잡구간, 최고))

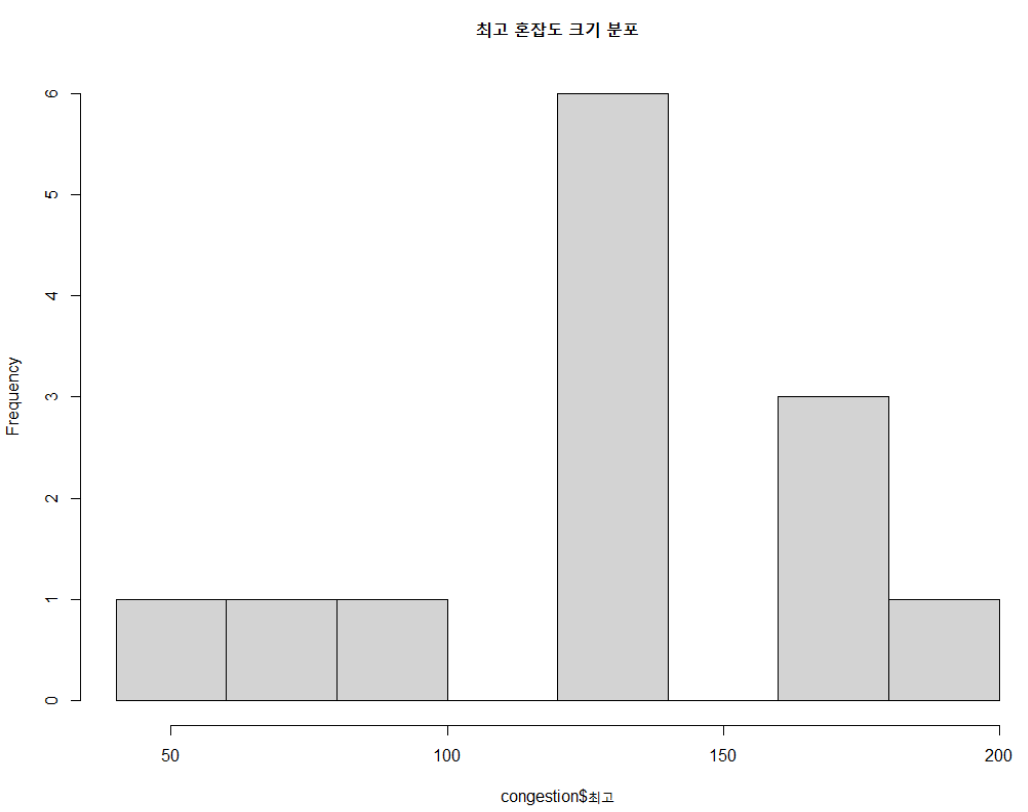
congestion



1. 분석

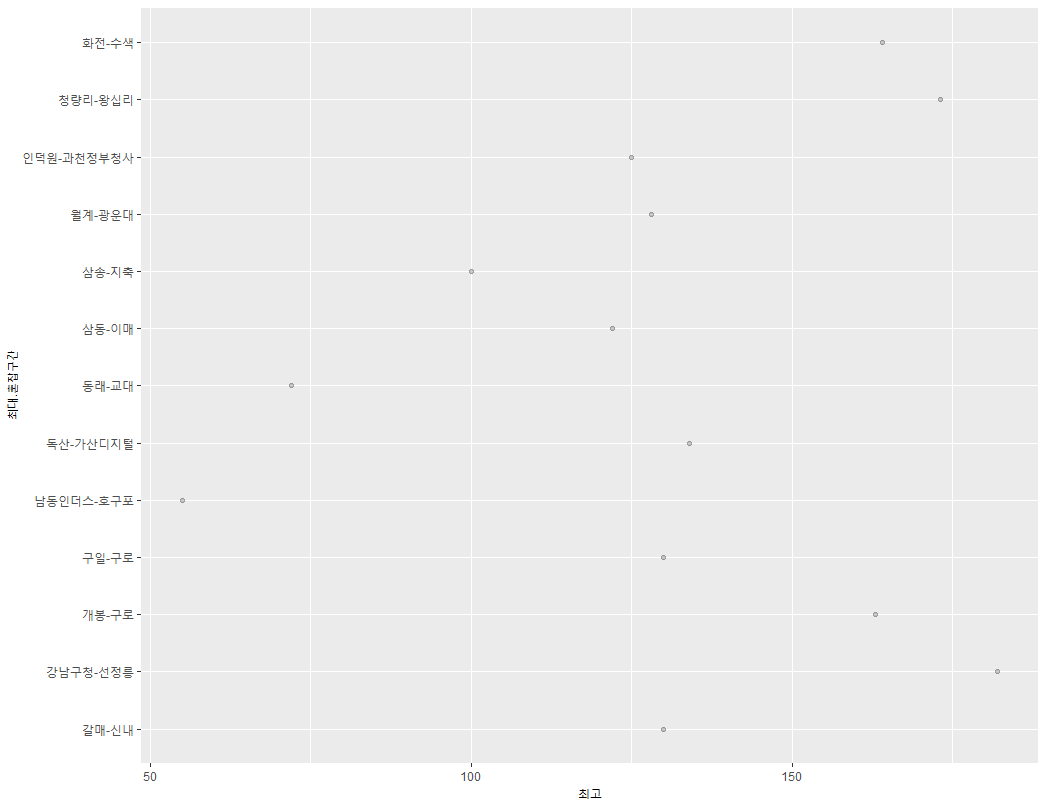
각 구간별 최고 혼잡도 크기를 비교해보니 모두 크기가 다른 것을 알 수 있다.

> hist(congestion$최고, main="최고 혼잡도 크기 분포")



어느 구간이 가장 혼잡한지 보기에는 어려움이 있어 ggplot함수를 사용하여 구간별 차이를 나타냈다.

> congestion %>% ggplot(aes(x=최고, y=최대.혼잡구간))+geom\_point(alpha=0.2)



수도권 혼잡 구간에 따른 최고 혼잡도를 비교해보니 지하철 역별로 세밀하게 혼잡도 정보를 알고싶어져 **서울교통공사 지하철 혼잡도 정보**를 수집하여 분석하였는데 이때, 너무 정보가 많아 출, 퇴근 가장 혼잡한 시간인 8시, 18시의 지하철 '하선'만 조사하였다

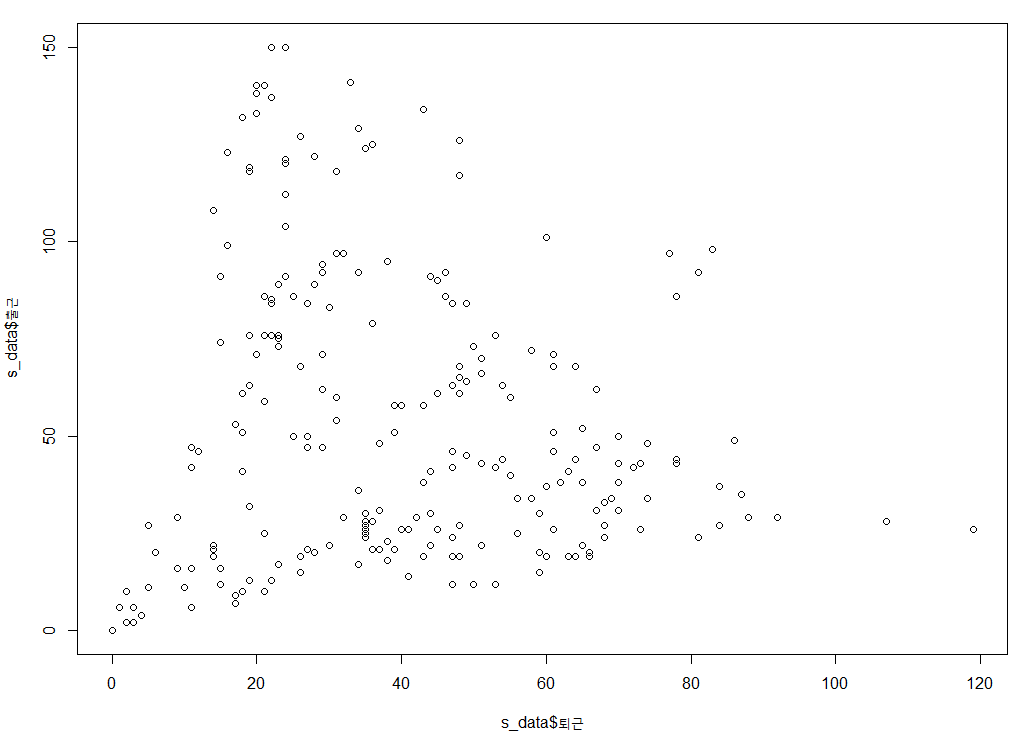
s\_data <- read.csv(file="C:/Users/User/Documents/서울교통공사\_지하철혼잡도정보\_20211231.csv", header=T, fileEncoding = "euc-kr")

s\_data <- subset(s\_data, 구분=='하선', select = -c(구분))

s\_data

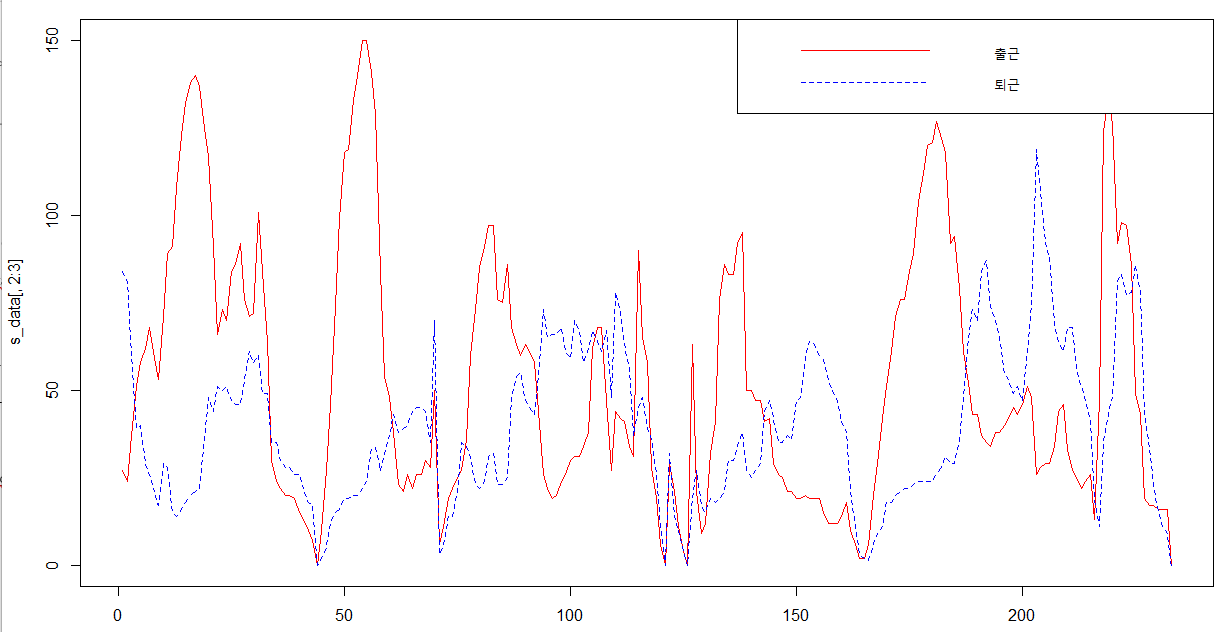
plot(s\_data$퇴근, s\_data$출근)

출근시간의 혼잡도가 퇴근시간의 혼잡도보다 높은 것을 알 수 있다.

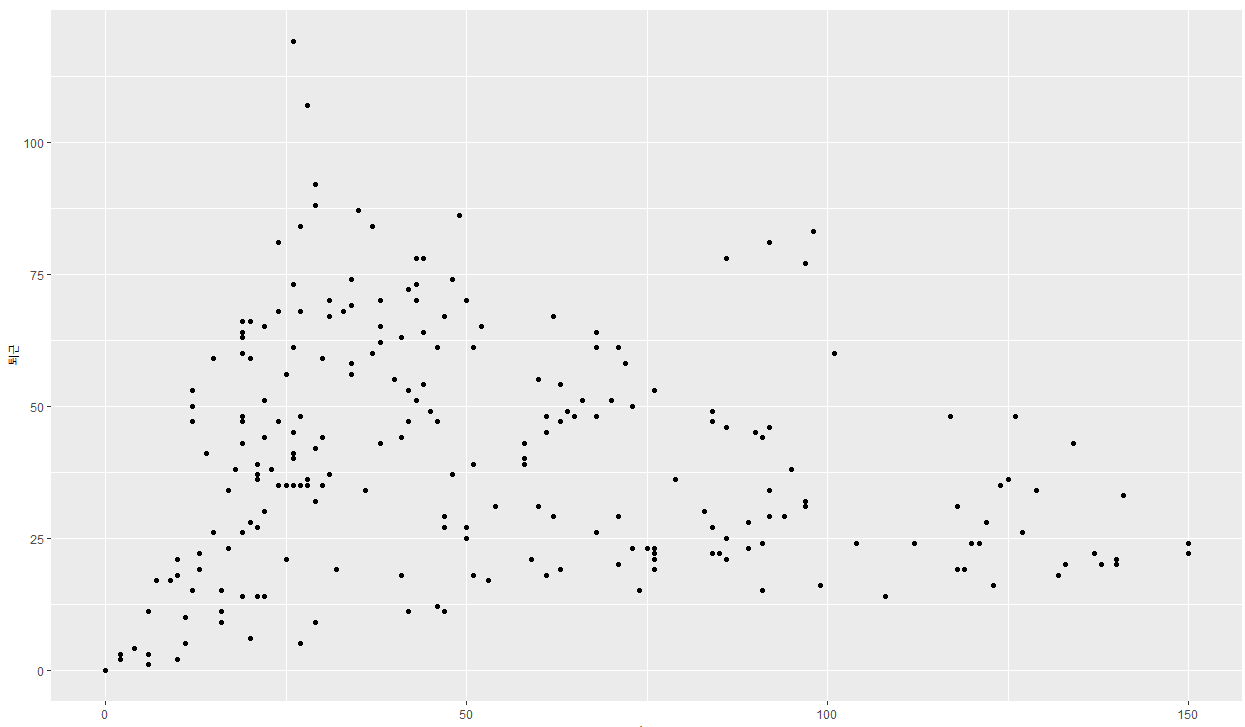


다른 방식

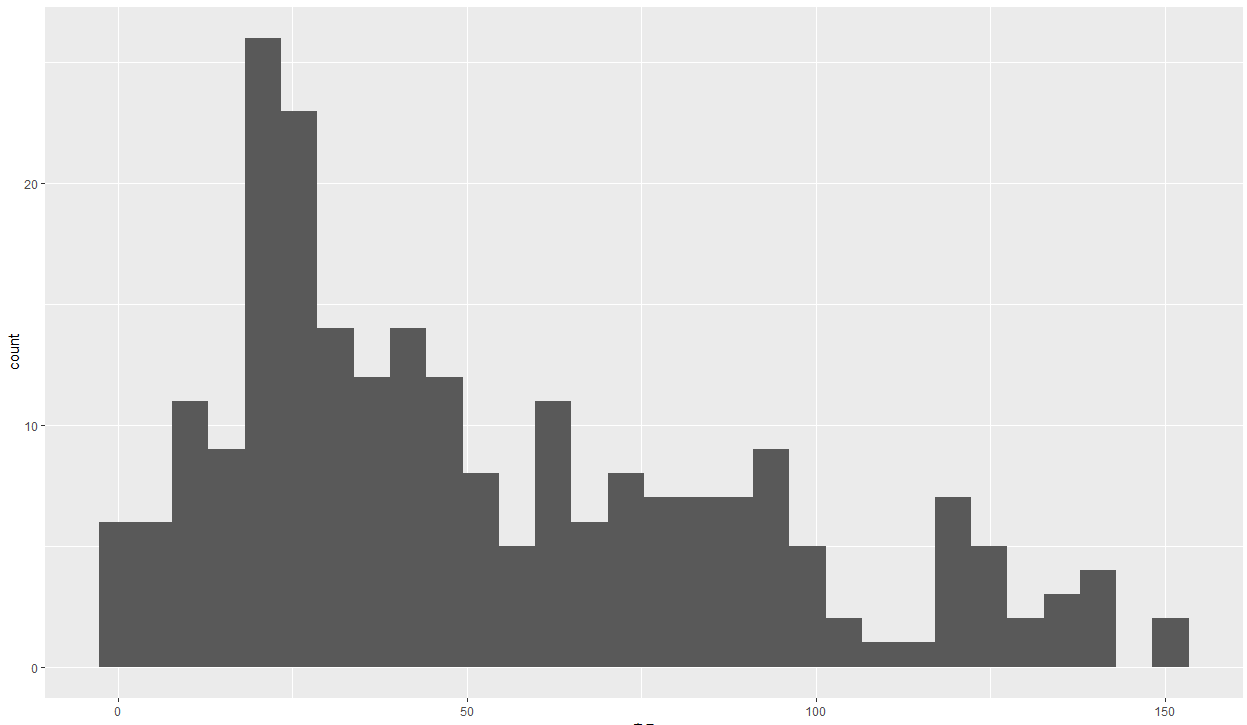
matplot(s\_data[, 3:4], type="l", col = c("red", "blue"))

legend("topright", names(s\_data)[3:4], lty=c(1,2), col = c("red", "blue"))

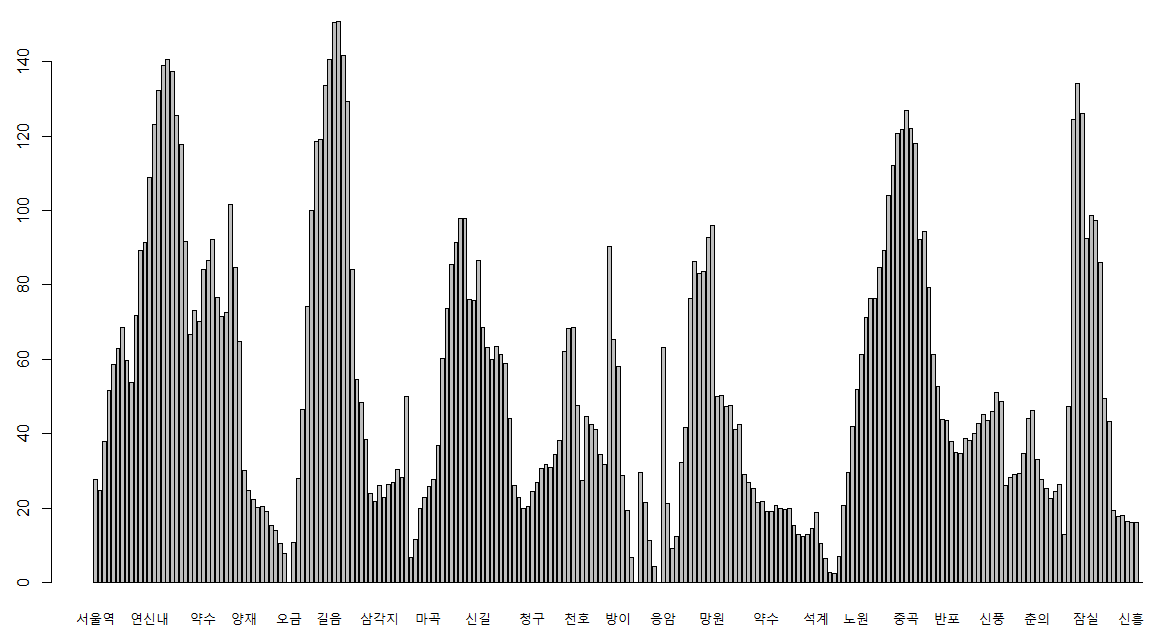
ggplot(s\_data, aes(x=출근, y=퇴근))+geom\_point()



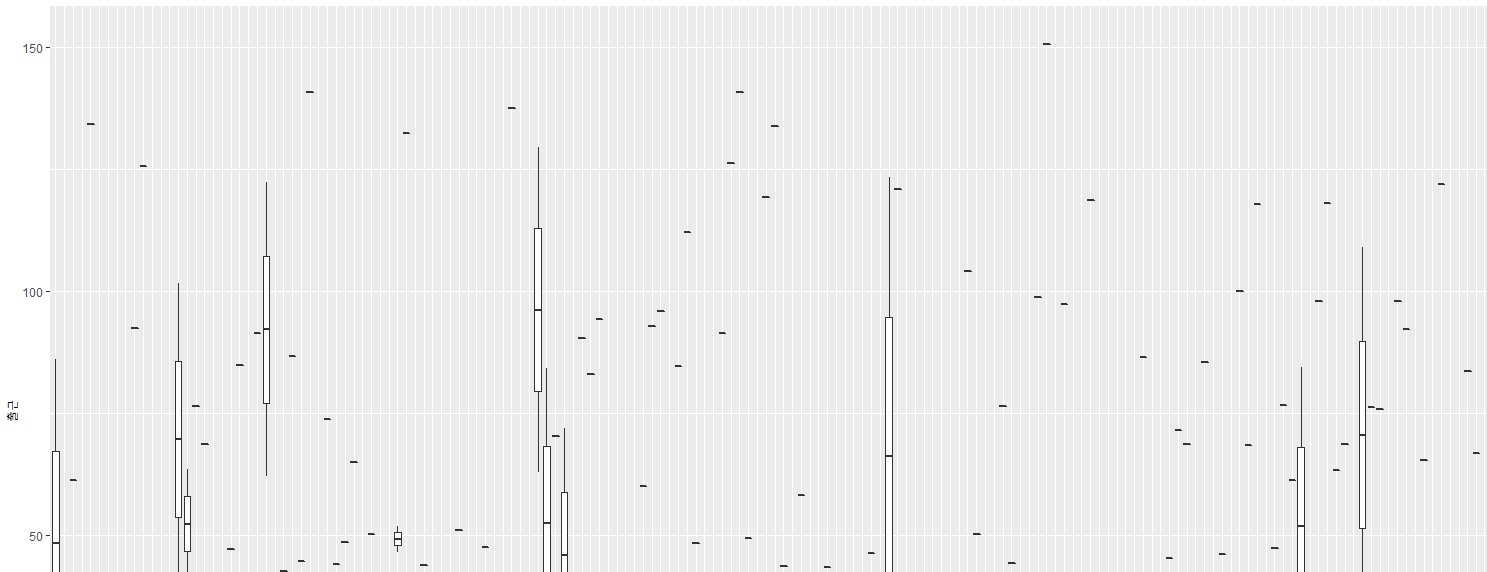
출근 시간에 더 혼잡하므로 평균 출근 시간인 8시의 역 별 혼잡도를 비교분석

ggplot(s\_data, aes(출근))+geom\_histogram()

barplot(s\_data$출근, names.arg=s\_data$역명)

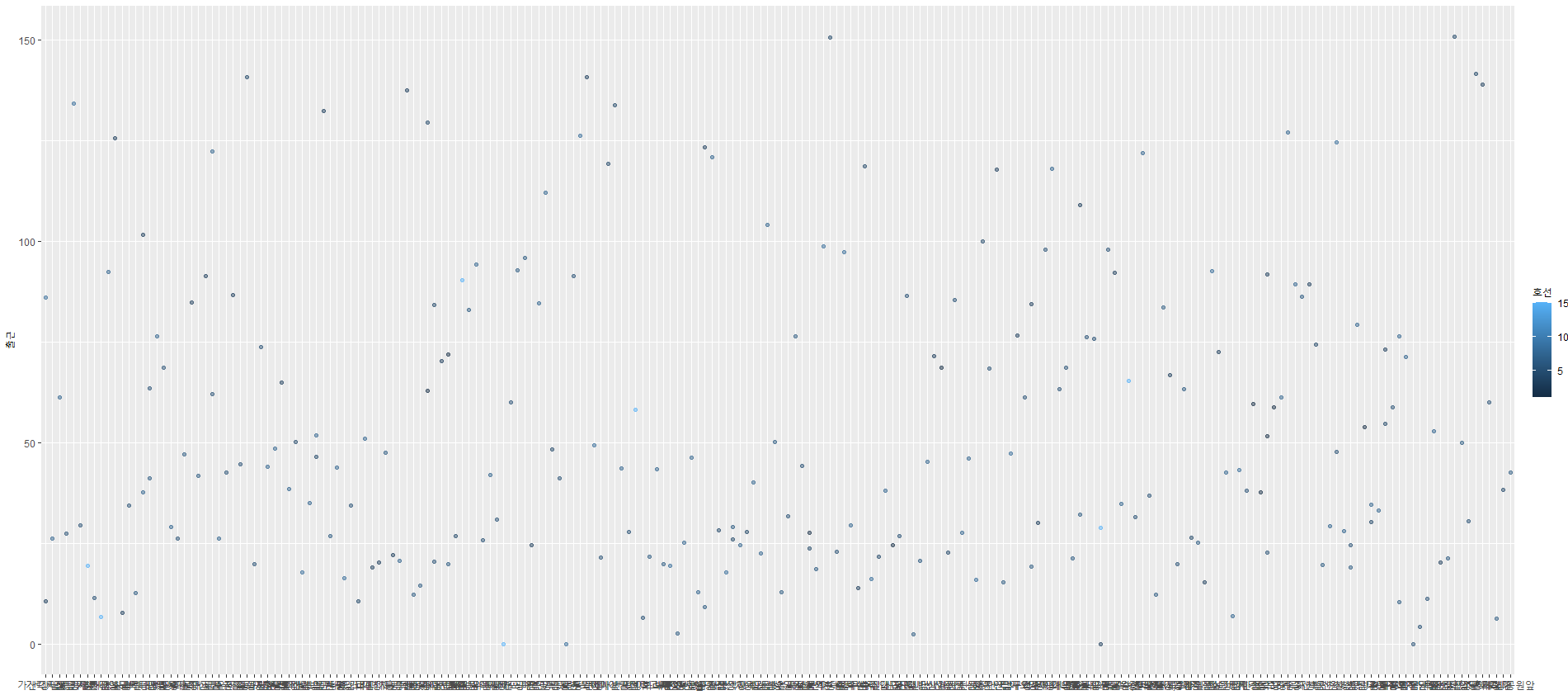


ggplot(s\_data, aes(역명, 출근))+geom\_boxplot()



호선 별 비교

ggplot(s\_data, aes(x=역명, y=출근, col=호선))+geom\_point(alpha=0.5)+geom\_smooth()



ggplot(s\_data, aes(x=역명, y=출근))+geom\_point(alpha=0.5)+facet\_wrap(~호선)