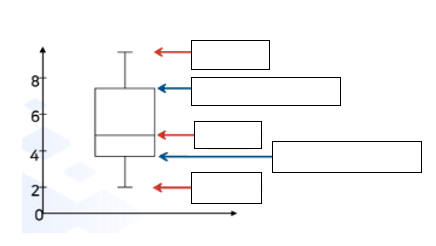
**HW2 (Due date : July.13 PM 19:59, upload in web)**

**Submit the homework named as HW2\_이름.doc, HW2\_이름.R (코드도 첨부)**

**문제**

1. **직교좌표의 평면에 관측점을 찍어 만든 통계 그래프로, 데이터 내 두 변수 간의 관계를 분석할 때 사용하는 시각화 기법은? 답 : 2번**
2. **histogram (2) scatter plot (3) bar plot (4) box plot**
3. **데이터 내 어떠한 단일 변수에 대해서 구간별 빈도수를 나타내는 그래프로, 해당 변수의 분포 알 수 있는 시각화 기법은? 답 : 4번**
4. **histogram (2) scatter plot (3) bar plot (4) box plot**
5. **아래 그림은 Box plot이다. Box plot의 구성 요소에 알맞게 빈칸을 채워 넣어라.**



최소값

중앙값

일사분위수(Q1, 25%)

삼사분위수(Q3, 75%)

최대치

1. **괄호 안을 채워 넣어라.**
2. **Box plot에서 Q1-1.5QR 값보다 작은 값을 갖거나, Q3+1.5QR 값보다 큰 값을 가진 데이터는 ( 이상치 )로 볼 수 있다.**
3. **IQR(quantile range)의 계산식: ( Q3 – Q1 )**

**실습**

**데이터설명 : U.S SAT scores by state for 2010 (SAT\_2010.csv)**

A data.frame with 50 rows and 9 variables.

state : a factor with levels for each state

expenditure : average expenditure per student (in each state)

pupil\_teacher\_ratio : pupil to teacher ratio in that state

salary : teacher salary (in 2010 US $)

read : state average Reading SAT score

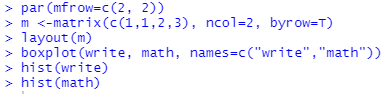
math state average Math SAT score

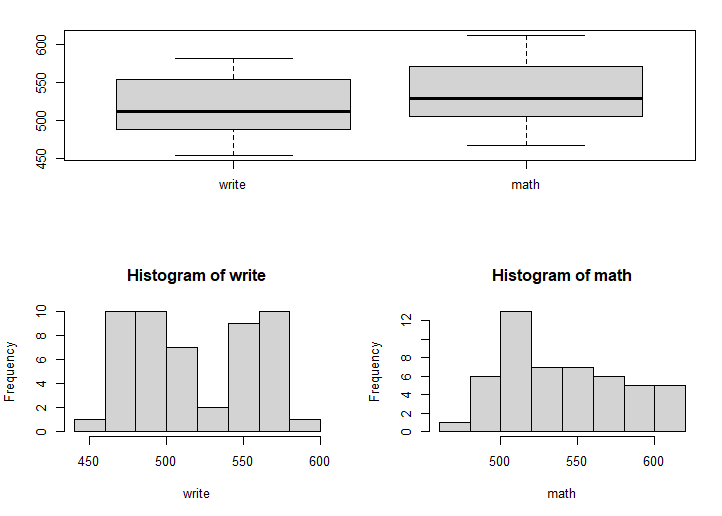
write state average Writing SAT score

total state average Total SAT score

sat\_pct percent of students taking SAT in that state

1. **상기 데이터에 대해 분석을 진행한다.**
   1. **“write”와 “math”에 대해 boxplot과 histogram을 그려라. 그린 그래프를 통해 각각의 분포를 설명하고 비교하라.**

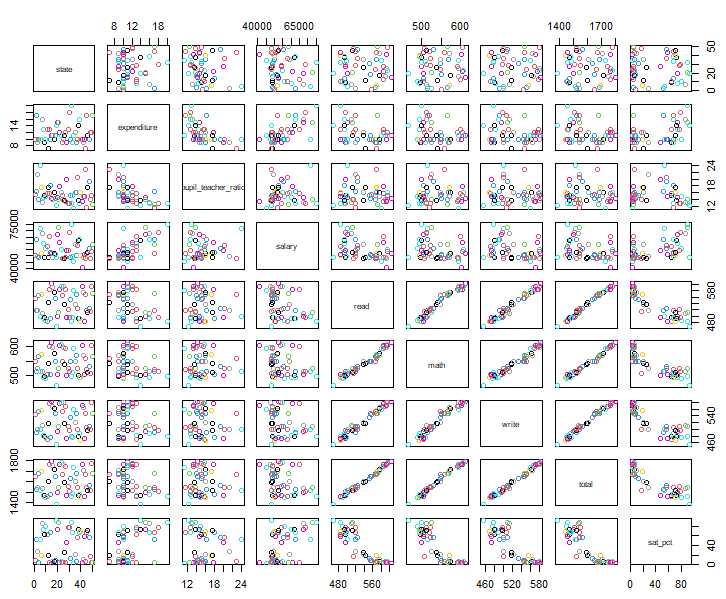




boxplot을 통해 비교해 볼 때, write와 math의 전체적인 분포는 write보다 math의 값이 큰 값을 가지고 있고, histogram을 통해 비교해 볼 때,

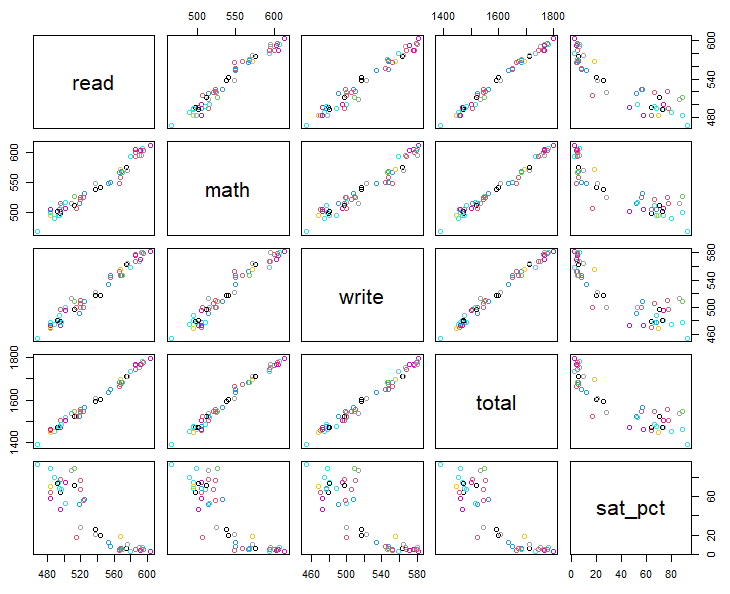
* 1. **산점도를 그려보고 “total” 변수와 가장 의미 있다고 생각하는 변수를 제시하고 설명해라**





데이터에서 상관관계를 볼 수 있는 산점도 그래프를 모두 그리고, 유의미해 보이는 그래프만 다시 추려서 출력해보면,

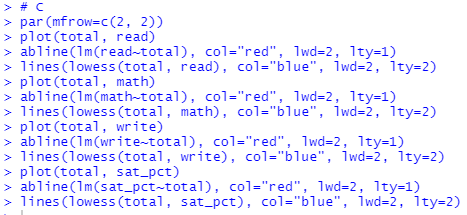


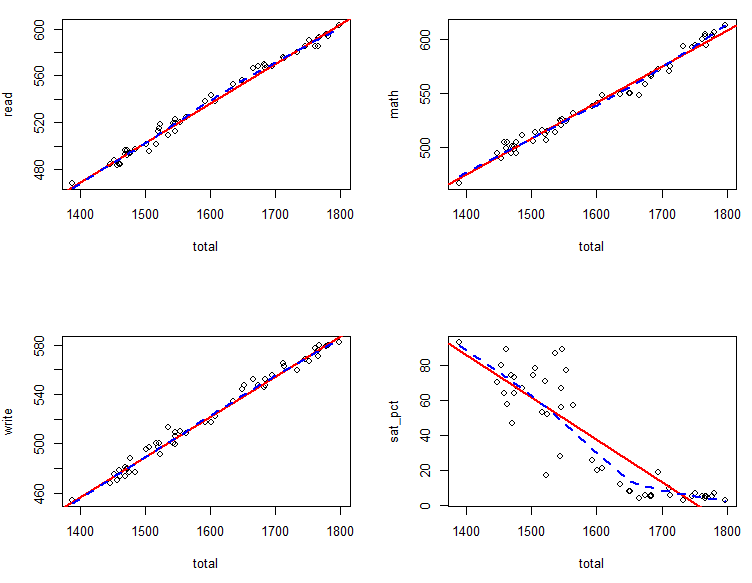


total과 read, math, write의 경우 비례하는 관계를 보여줍니다. 그리고 read와 math, write는 서로 각각 비례하면서 서로의 값의 의미가 중복되는 경향이 있음을 보여줍니다.

그리고 total과 sat\_pct의 경우, 어느정도 반비례 하는 모양을 보여줍니다.

* 1. **B에서 선택한 변수와 total 변수 간의 산점도를 그리고 이를 바탕으로 최적적합함수를 추정하여 그래프로 그려라. (선형, 비선형)**





**데이터 설명 : BTS, zoom, 라면, 코로나에 대한 검색결과**

**search.csv**

일: 날짜 (YYYY-MM-DD)

라면 : 주어진 기간 가장 큰 검색량을 100으로 한다. 그 기준 날짜에 라면의 검색량.

Zoom : 주어진 기간 가장 큰 검색량을 100으로 한다. 그 기준 날짜에 zoom의 검색량.

코로나 : 주어진 기간 가장 큰 검색량을 100으로 한다. 그 기준 날짜에 코로나의 검색량.

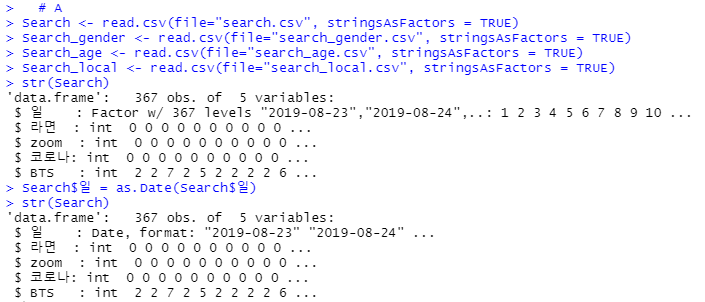
BTS : 주어진 기간 가장 큰 검색량을 100으로 한다. 그 기준 날짜에 BTS의 검색량.

**search\_gender.csv**

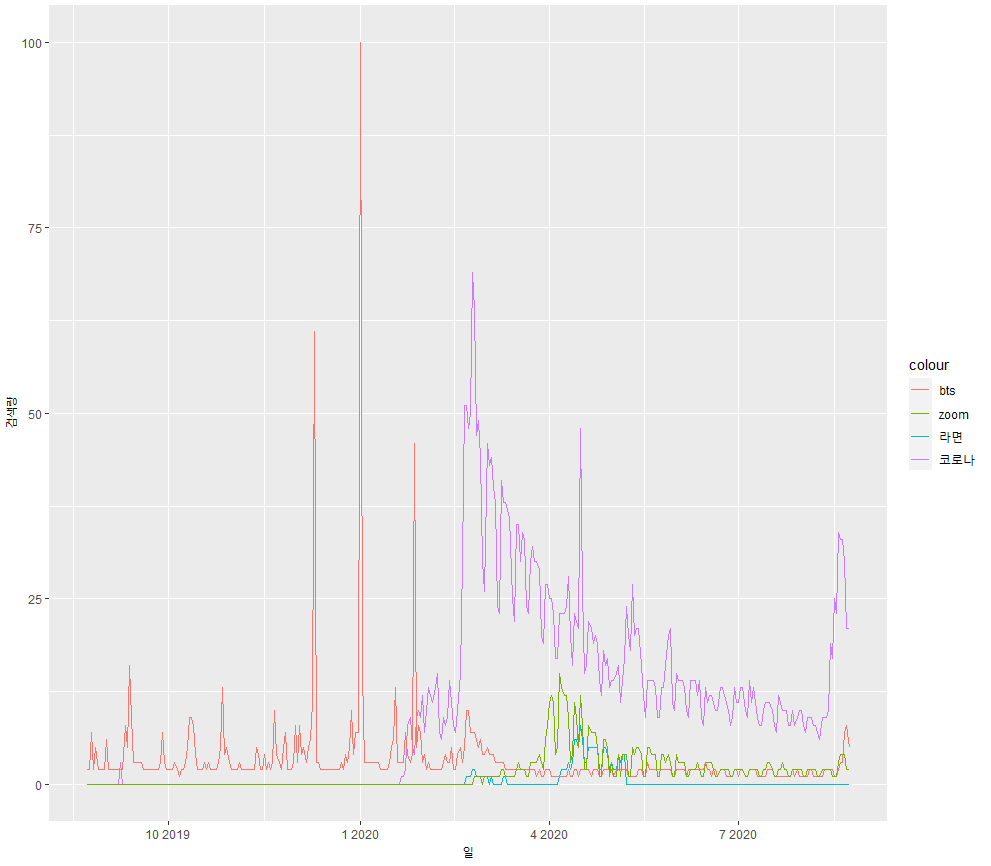
**search\_age.csv**

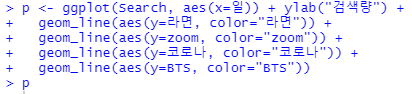
**search\_local.csv**

1. **상기 데이터에 대해 분석을 진행한다. (ggplot 사용)**
   1. **주어진 데이터 search.csv, search\_gender.csv, search\_age.csv, search\_local.csv를 불러들여라 (search.csv를 data로 읽었다 하자. data$일 = as.Date(data$일) 을 적용하라.)**

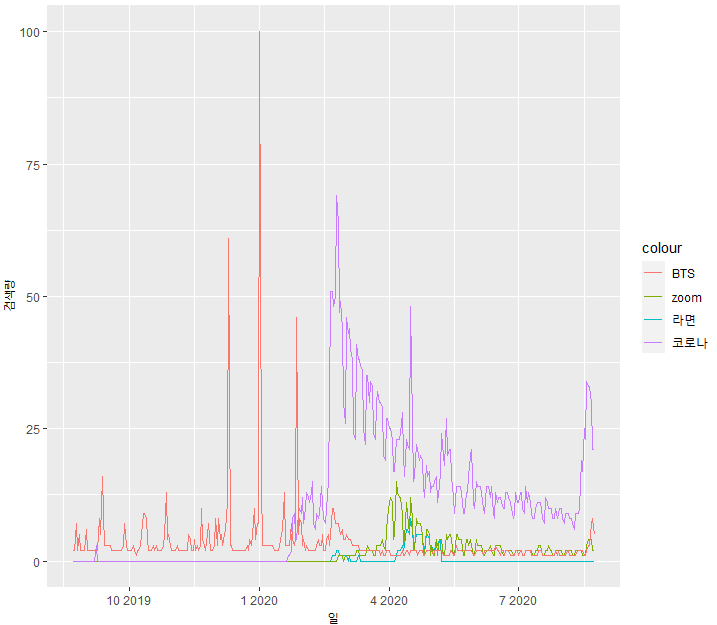


* 1. **검색량.csv 데이터를 geom\_line을 통해 다음과 같이 표현후, 일자별 검색량 변화를 설명하라.**

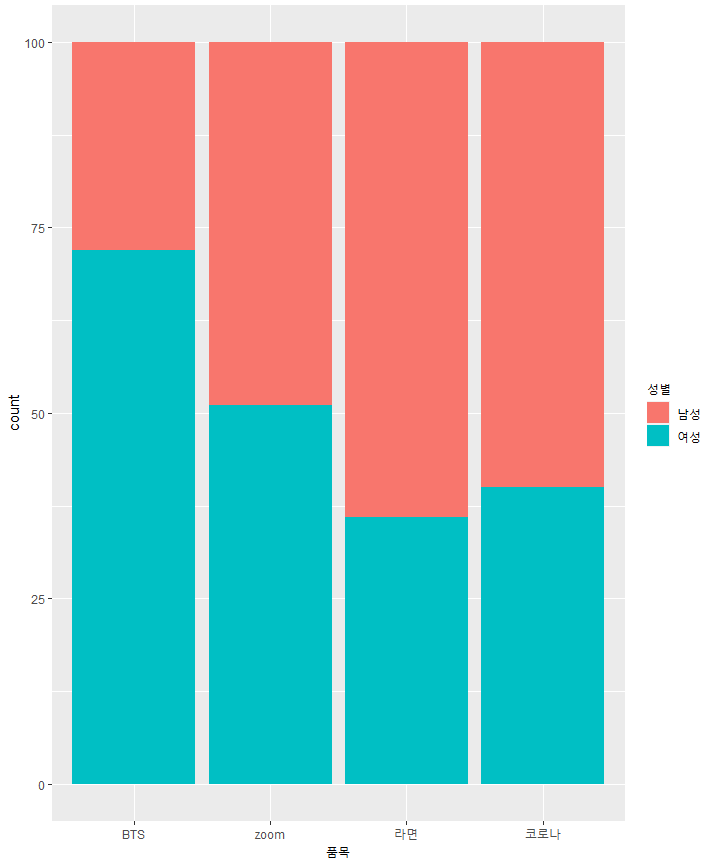
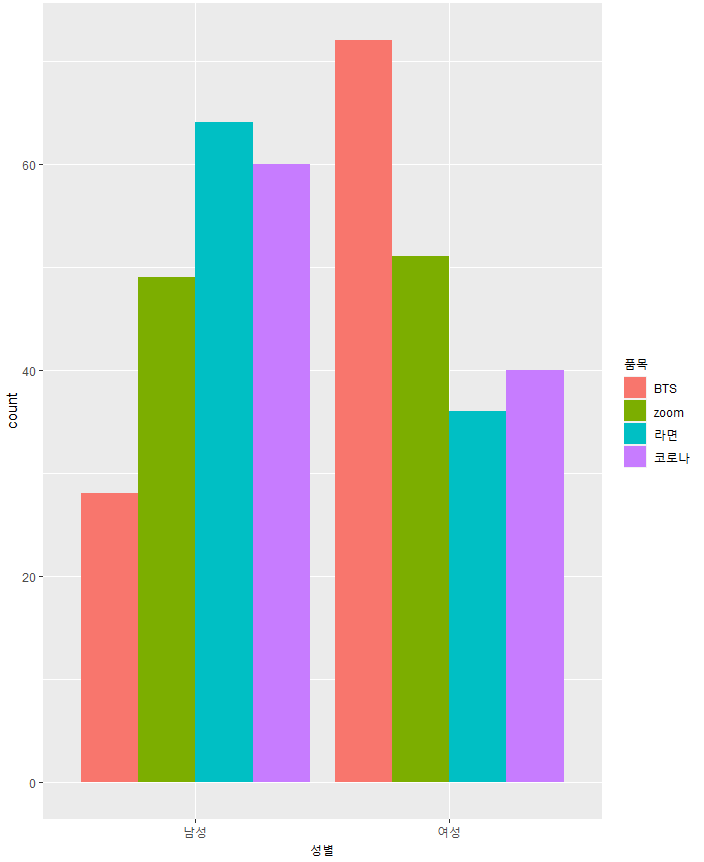




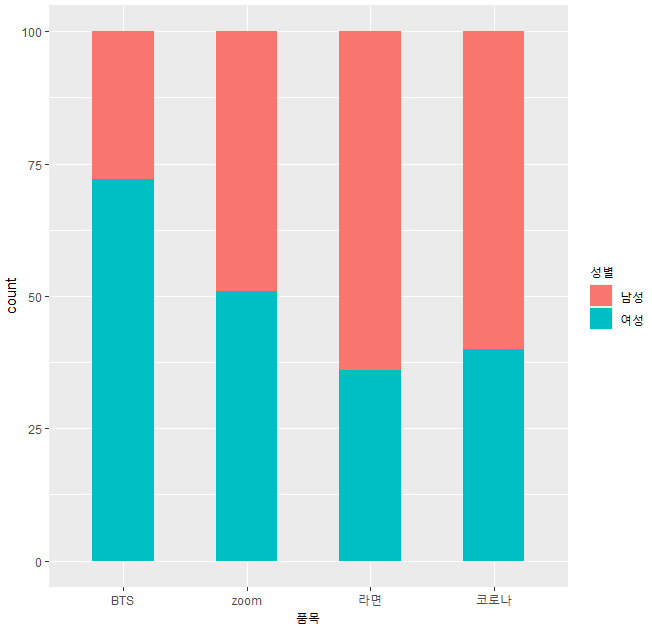
카테고리 별 검색량 변화량을 그래프를 보고 알 수 있으며, BTS의 경우가 검색량이 최고점을 찍었습니다. 라면의 경우가 검색량의 최고점이 가장 낮습니다. 2020년 2월 중순을 기준으로 코로나에 대한 검색량이 급증하였고, 그 시기를 기점으로 BTS에 대한 검색량이 대폭 줄었습니다.



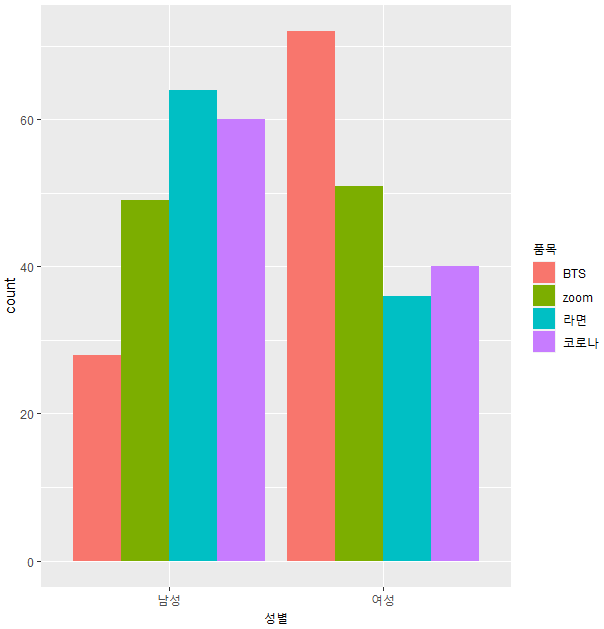
* 1. **검색량\_성별.csv데이터를 다음과 같이 bar chart로 표현하고, 성별 및 품목별 검색량 차이를 비교하라.**

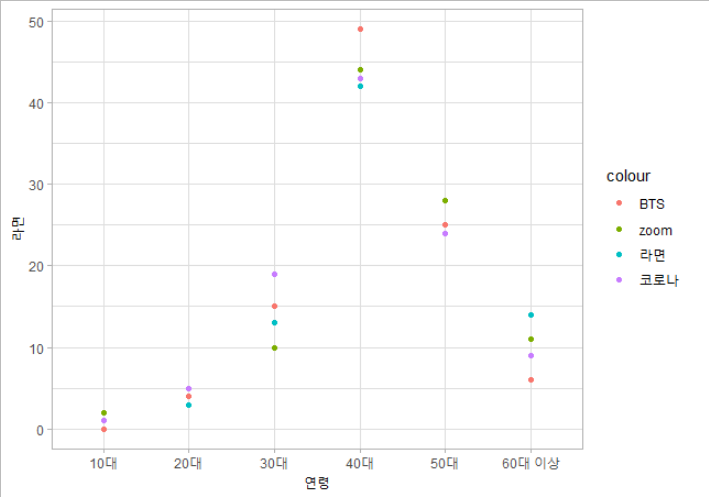


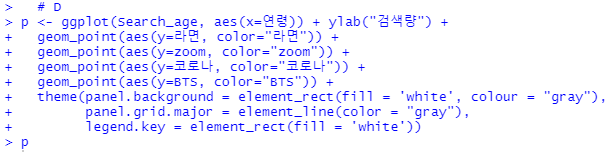


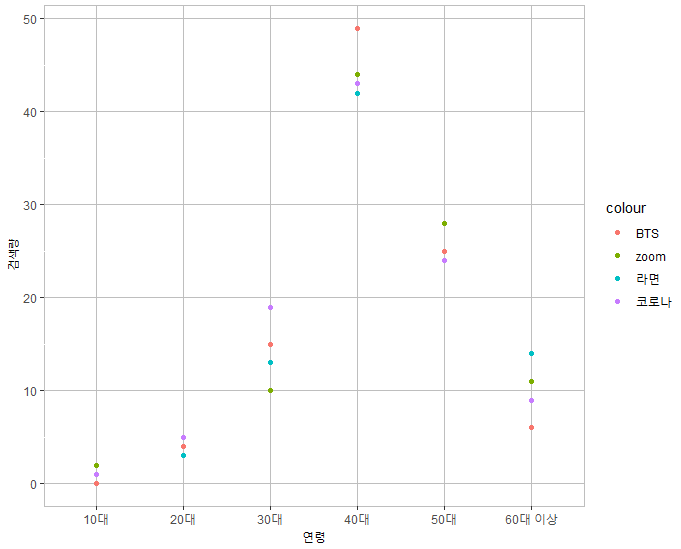




* 1. **검색량\_연령.csv데이터를 다음과 같이 scatter Plot으로 표현해라**

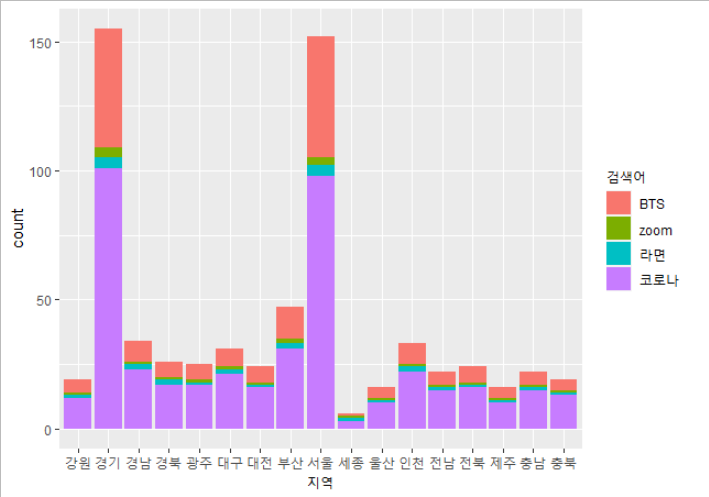


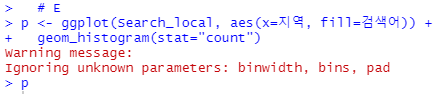


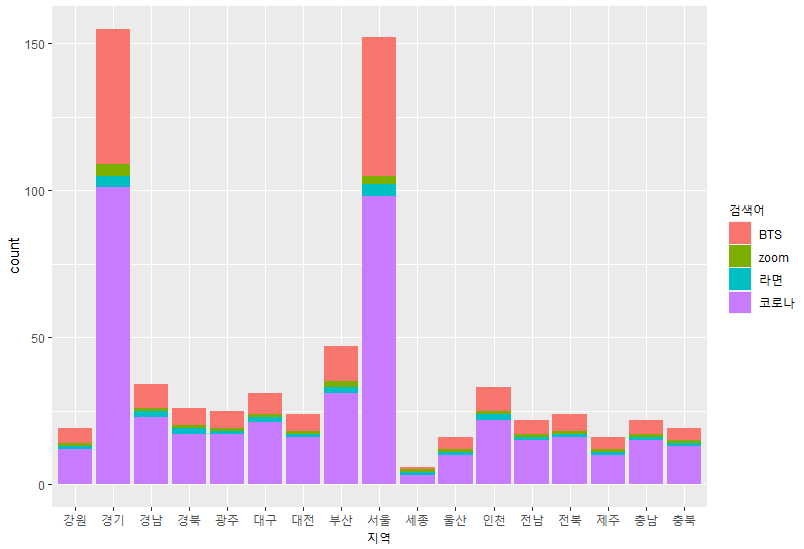


* 1. **검색량\_지역.csv데이터를 다음과 같이 Histogram으로 표현해라.**

**(geom\_histogram(stat=”count”)사용)**







* 1. **위의 표현된 결과를 통해 각각 BTS, zoom, 라면, 코로나의 지역별 검색량을 비교하라.**

지역 전체적으로 ‘코로나’라는 검색어가 빈도수가 가장 많고, 그 다음으로 BTS가 많습니다. 서울과 경기의 경우 다른 지역들에 비해 검색량이 압도적으로 많은 것으로 보이고, 세종시가 가장 적습니다. ‘zoom’과 ‘라면’의 경우 검색량이 나머지 두 개 보다 훨씬 적습니다.