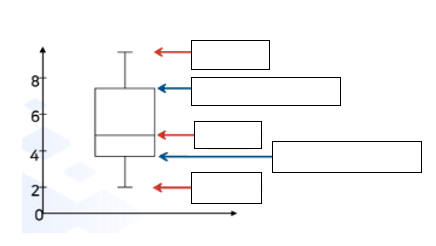
**HW2 (Due date : July.13 PM 19:59, upload in web)**

**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? 2**
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? 1**
4. **histogram (2) scatter plot (3) bar plot (4) box plot**
5. **아래 그림은 Box plot이다. Box plot의 구성 요소에 알맞게 빈칸을 채워 넣어라.**



**최대치**

**삼사분위수(Q3, 75%)**

**중앙값**

**일사분위수(Q1, 25%)**

**최소값**

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

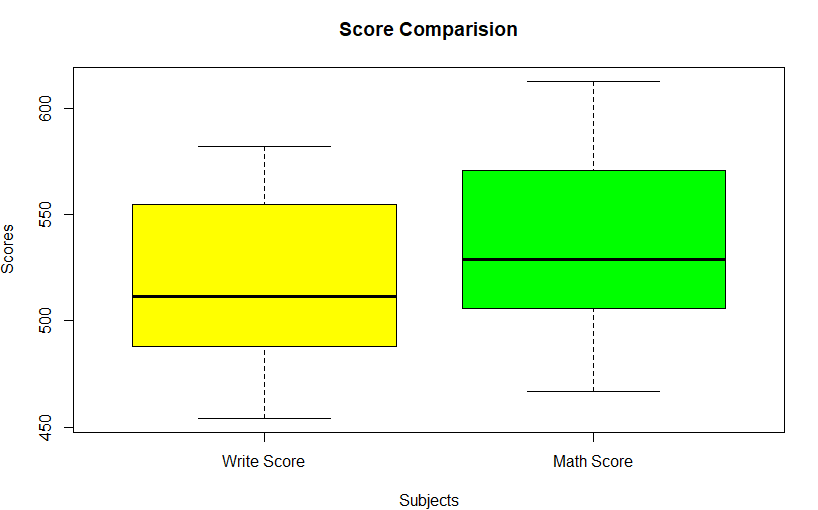
math state average Math SAT score

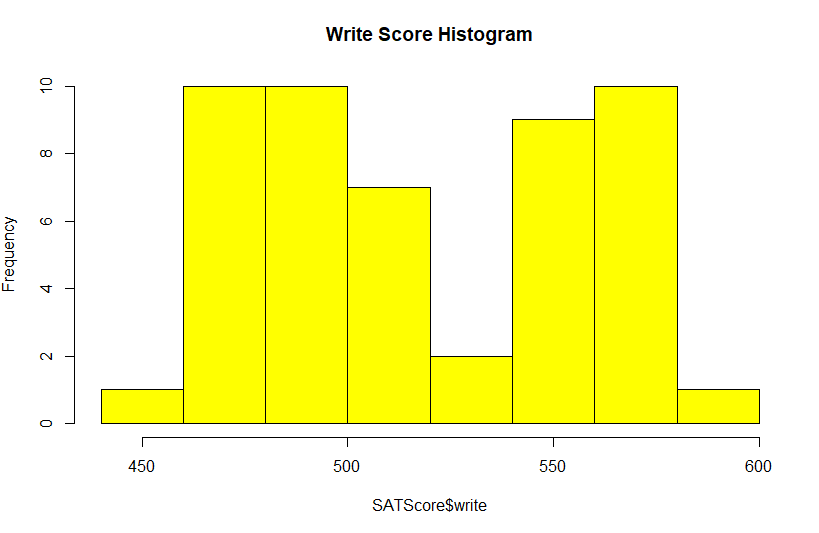
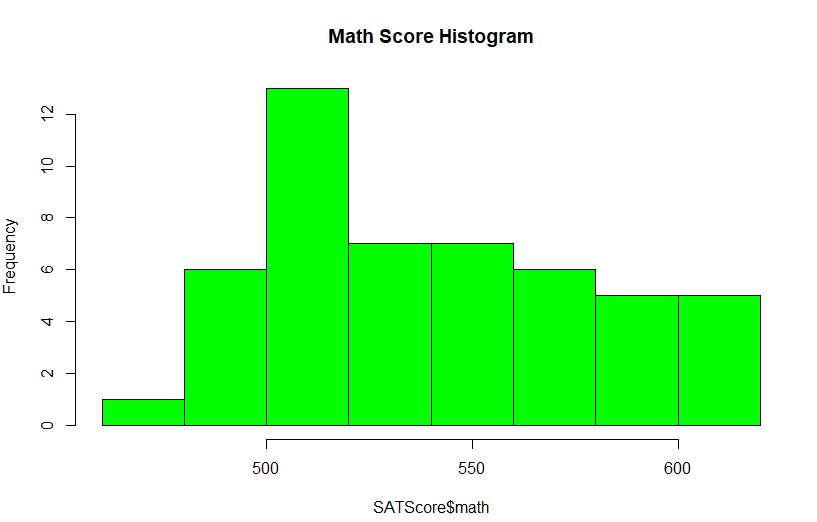
write state average Writing SAT score

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.**

****

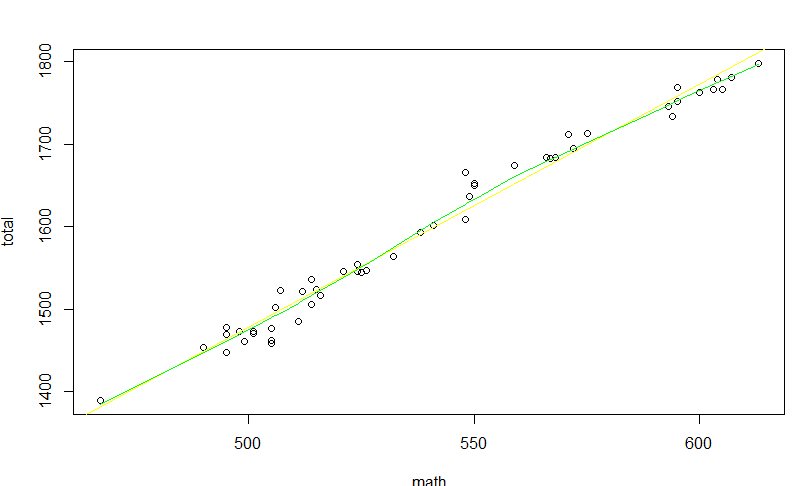
****

**475점에서 500점대와 550~ 575점대에 많은 점수가 몰려있는 것을 히스토그램으로 확인할 수가 있다. boxplot을 통해 math의 중앙값이 write 중앙값보다 높다는 걸 확인할 수 있다. 뿐만 아니라 최소값과 최대값 일사 분위수와 삼사 분위수가 크므로 write점수가 더 높다고 이야기 할 수 있다.**

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

육안으로 선형회귀를 가장 잘 나타나는 read 변수가 가장 의미 있는 거 같습니다.

* 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.)**

**search = read.csv("search.csv")**

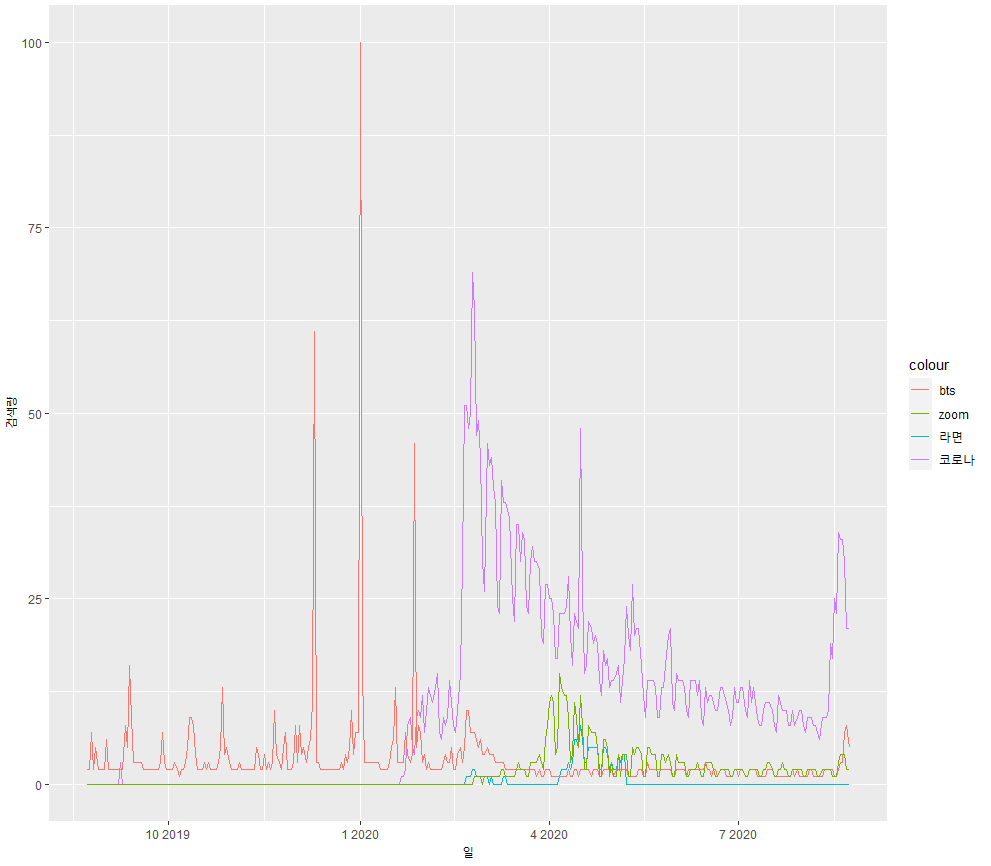
**search\_age = read.csv("search\_age.csv")**

**search\_gender = read.csv("search\_gender.csv")**

**search\_local = read.csv("search\_local.csv")**

**search\_df$일=as.Date(search\_df$일)**

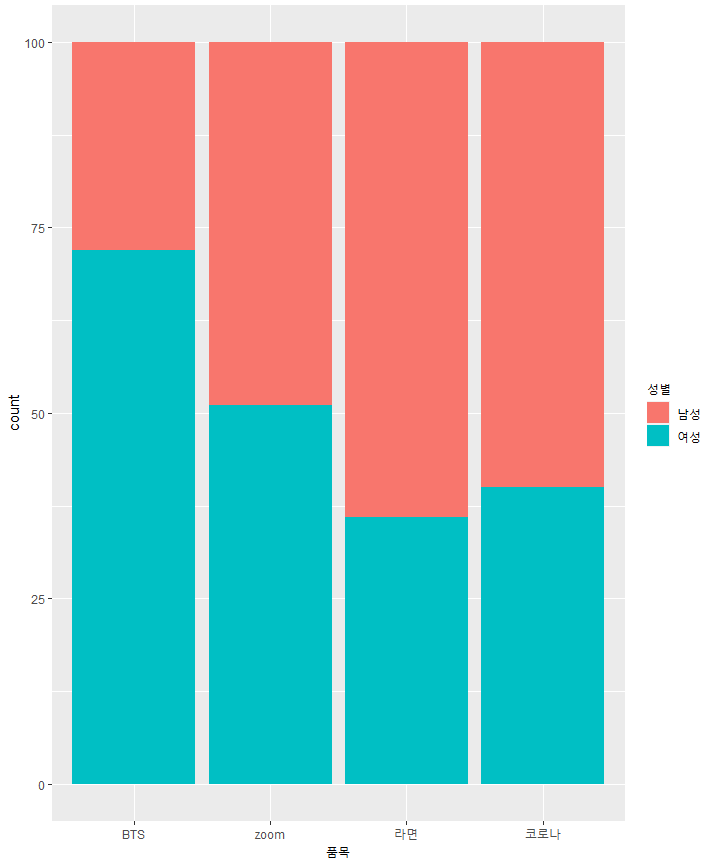
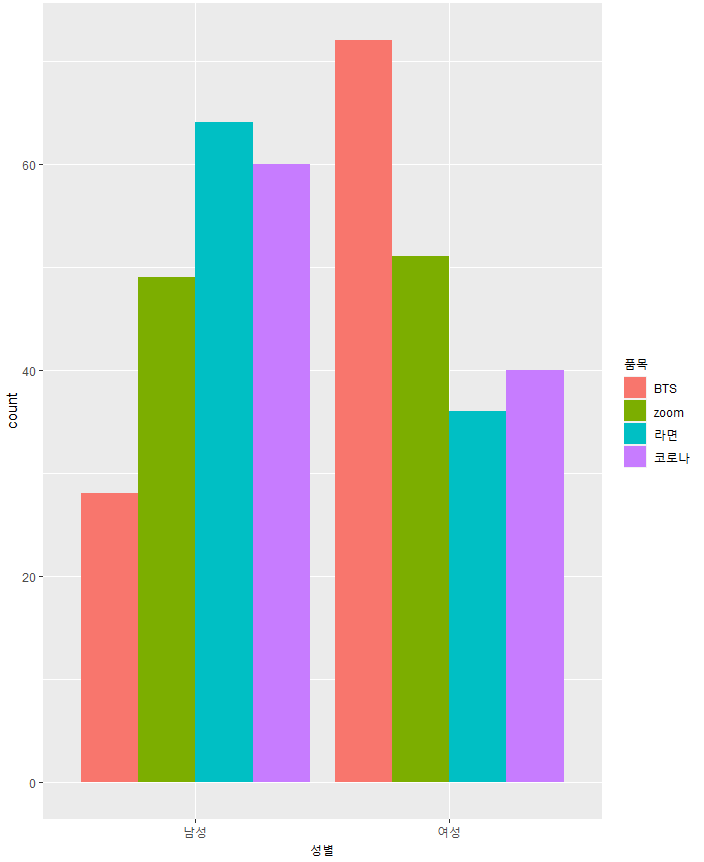
* 1. **검색량.csv 데이터를 geom\_line을 통해 다음과 같이 표현후, 일자별 검색량 변화를 설명하라.**



**2020년까지는 BTS 관련 검색이 아주 많았고 코로나의 대유행이 시작되면서 검색량이 많아졌다.**

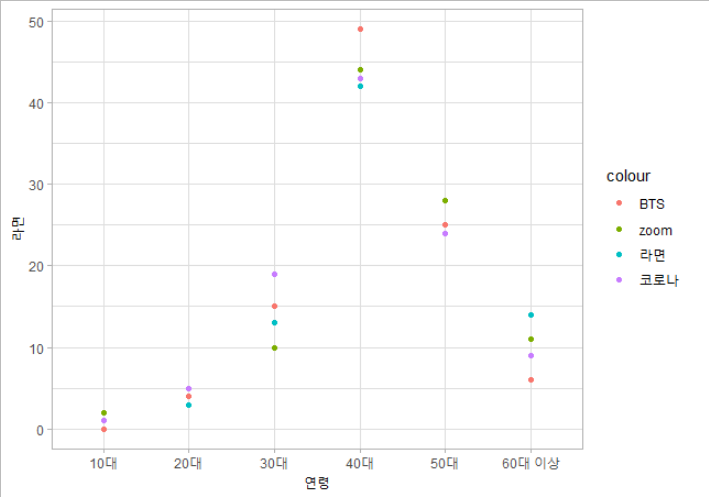
**코로나가 지속되는 와중에 라면과 zoom에 대한 검색량이 늘었고 상관관계가 있다는 것을 유추할 수 있다.**

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

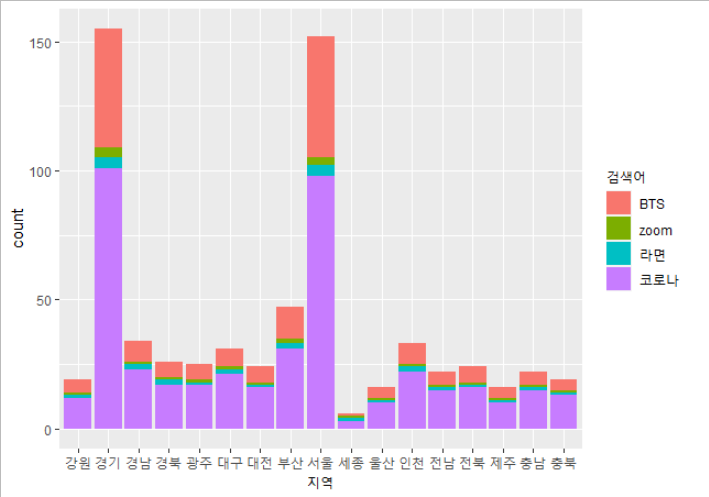
조금 뻔한 결과가 나오는 듯 하다. BTS같은 경우에 당연히 여성이 검색량이 많을거라는 예상을 할 수 있었고 실제로 데이터도 그렇다. zoom은 남녀 비슷한 검색량을 보인다. 이게 아마도 zoom은 남녀차별없이 과제, 면접과 먹고 사는일 때문에 아닌가 싶다. 이번 데이터에서 제일 흥미로운 점은 라면에 대한 검색량이 남자들이 더 많이 보였다는 것이다. 실제로 라면은 남자들이 여자들에 비해 얼만큼 더 많이 섭취하고 찾는지에 대한 데이터 분석을 하고 싶어진다.

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



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

**(get\_histogram(stat=”count”)사용)**

**ggplot(data=search\_local,aes(x=지역, fill=검색어))+geom\_histogram(stat="count")**

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

**인구수가 집중되어 있는 서울과 경기는 확연하게 검색량이 많다. 1순위가 코로나이고 2순위는 BTS인걸 알 수 있다. 아무리 인구수가 차이가 많이 난다고 해도 인터넷을 사용하는 층이 서울과 경기가 지방에 비해 압도적으로 많은 것이 놀랍다. 유행과 정보의 속도가 수도권이 현저히 빠르다는 걸 데이터로 체감할 수 있다.**