**9/2 과제: 데이터 탐색 및 t-test**

**데이터 설명: World Happiness Report for 2019 (WHR\_2019.csv)**

A data.frame with 50 rows and 9 variables.

Overall Rank : 나라별 순위

Country or region : 나라 혹은 지역

Score : score of happiness

GDP per capita : 1인당 국민 소득

Social support : 사회적 지원

Healthy life expectancy : 건강 평균 수명

Freedom to make life choices : 자유

Generosity : 관대함, 너그러움

Perceptions of corruption: 정치 부패에 대한 인식

1. **데이터 다루기**

**1.1 data1, data2, data3을 하나의 데이터 프레임으로 결합시켜라. (merge, rbind 함수 사용)**

**1.2 데이터 정렬**

**①1.1에서 얻은 데이터를 변수 Overall rank(Score을 사용할 때에는 decreasing=T)로 정렬을 하고 이 때 상위 5위의 국가를 모두 나열해라.**

답:Finland, Denmark, Norway, Iceland, Netherlands

**②1.1에서 얻은 데이터를 변수 GDP per capita로 먼저 정렬을 하고 그 다음 변수 Generosity로 정렬시켜라. 정렬 후, 나라의 순서가 바뀌는데 이 때 하위 5위의 국가를 모두 나열해라.**

답:Somalia, Central African Republic, Burundi, Liberia, Congo

**1.3 1.1에서 얻은 데이터에서 *행복 지수가 점수가 높은 순서대로 1~100위까지* 데이터를 추출해서 저장해라.**

**1.4 1.3에서 얻은 데이터에서 변수 Overall rank를 제외하고 df\_whr 이름의 데이터로 저장해라.**

1. **데이터의 기술 통계치 요약**

**(1번 문제에서 저장한 df\_whr 데이터 계속 사용하세요)**

**2.1 Score 변수의 분산과 평균값, 중앙값은?**

답:0.5439004, 6.07257, 6.0245

**2.2 Score 변수가 (평균-표준편자, 평균+표준편차) 즉, 1시그마 범위 내에 속하는 나라의 개수는?**

답:59개

**2.3 Score의 평균으로부터 가장 멀리 떨어진 Score값을 가진 나라는 어디인가?**

답:Finland

**데이터 설명: Heart Disease UCI (heart.csv)**

age : age in years

sex : (1 = male; 0 = female)

cp : chest pain type

trestbps : resting blood pressure (in mm Hg on admission to the hospital)

chol : serum cholestoral in mg/dl

fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg : resting electrocardiographic results

thalach : maximum heart rate achieved

exang : exercise induced angina (1 = yes; 0 = no)

oldpeak : ST depression induced by exercise relative to rest

slope : the slope of the peak exercise ST segment

ca : number of major vessels (0-3) colored by flourosopy

thal : 3 = normal; 6 = fixed defect; 7 = reversable defect

target : have disease or not (1=yes, 0=no)

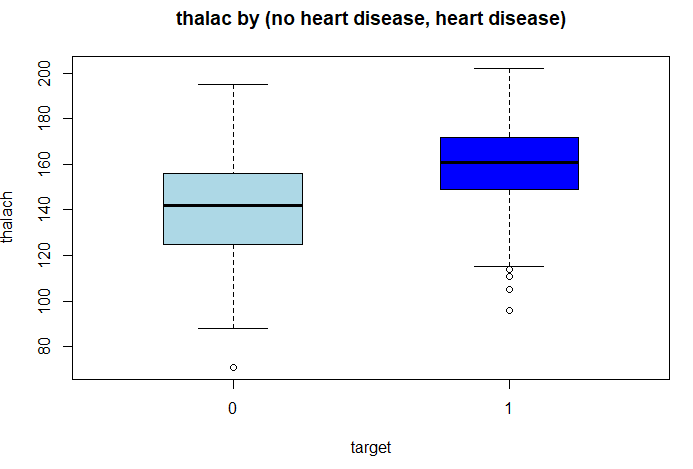
**3. 그래프를 이용한 데이터 탐색**

**3.1 나이별로(변수 age) 데이터 분포가 어떻게 되어 있는지 히스토그램을 그려서 확인해라(break=10). 어느 나이대에 사람들이 가장 많은가?**



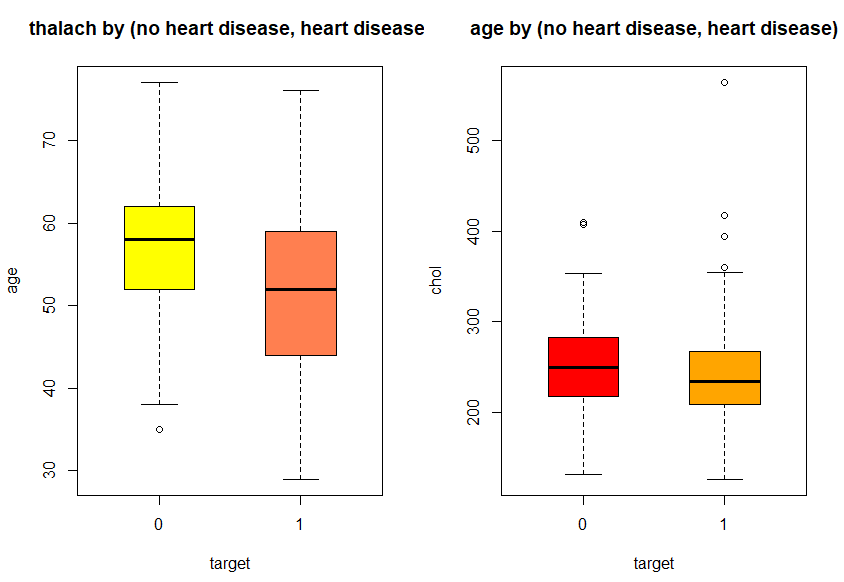
답 : 55~60대

**3.2 변수 thalach에 대해 변수 target을 x축으로 삼아 boxplot을 그리고, 분포 및 해당 결과에 대해 설명해라.**



heart disease에 걸리는 사람들이 최고 혈압 수치가 평균적으로 더 높다.

**추가 문제**(푸셔도 되고 안 푸셔도 됩니다)**:** 변수 thalac에 대해 변수 target을 x축으로 삼아 boxplot을 그리고, 분포 및 해당 결과에 대해 설명해라.

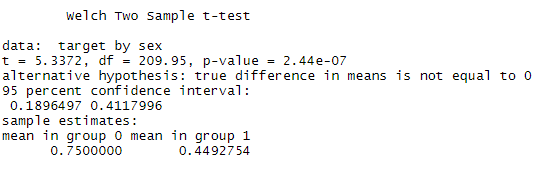


heart disease에 걸리는 사람들이 평균적으로 나이가 더 젊다.

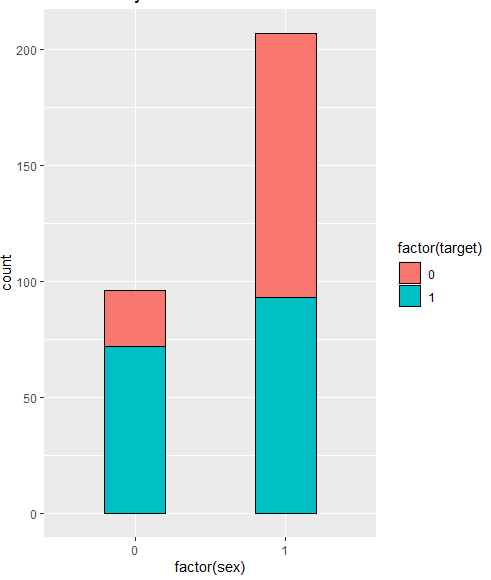
heart disease에 걸리는 사람들이 평균적으로 콜레스테롤 수치가 더 낮다.

**4. t-test**

**4.1 성별 별로 heart disease의 발병 여부에 대해서 t-test를 진행하고(정규 분포라 가정함) 해당결과에 대해서 설명해라. 막대 그래프로 이를 다시 설명해라(ggplot)**



p-vlaue값이 0.05 이하인 것으로 보아 성별이 heart disease에 유의한 영향을 끼친다는 것을 알 수 있다.



해당 막대 그래프를 보면 여성(여성이 0, 남성이 1) 중 heart disease에 걸린 사람의 비율이 훨씬 크다는 것을 확인할 수 있다 =>sex가 heart disease에 유의한 영향을 끼친다.