**HW4 (Due date : July.15 PM 19:59, upload in web)**

**Submit the homework named as HW4\_이름.doc, HW4\_이름.R (코드도 첨부)**

**문제**

1. **상관분석에 대한 설명으로 옳지 않은 것은? 3번**
2. **절대값이 0에 가까울수록 상관관계가 없다.**
3. **절대값이 1에 가까울수록 강한 상관성이 있다.**
4. **상관계수 값의 범위는 0~1사이이다.**
5. **상관계수를 통해 두 변수의 선형 관계를 확인할 수 있다.**
6. **회귀분석에서 결정 계수란 무엇인가?(SST와 SSR의 정의 설명 필요)**

**결정계수 은 전체제곱합(SST)에 대한 회귀제곱합(SSR)의 비율, 즉 모형으로 설명 할 수 있는 부분의 비율을 말한다.**

**SSR은 회귀식에 의해 설명되는 부분을 의미하고**

**SSE는 회귀식에 의해 설명되지 않는 부분(오차의 합)을 의미한다.**

**결정계수** **은 0<=<=1 사이 값이고, 이다.**

**은 1에 가까울수록 회귀식에 적합되는 부분이 높고, 0에 가까울수록 주어진 독립변수들에 의해 설명되는 부분이 없다고 할 수 있다.**

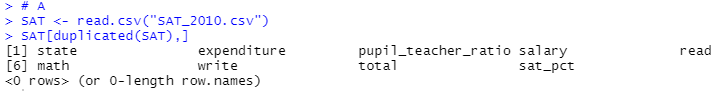
1. **괄호 안을 채워 넣어라.**

**( 텍스트마이닝 )이란 웹페이지, 이메일, 소셜네트워크 기록 등 전자문서 파일로부터 특정 연관성(동시적으로 빈도가 높은 단어 추출)을 분석하는 방법이다.**

**실습**

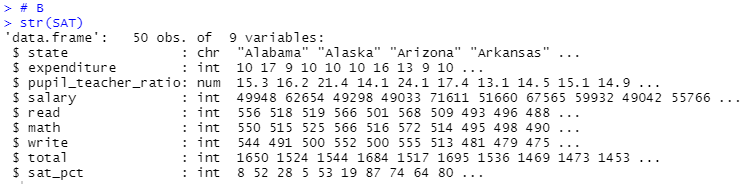
**데이터 설명 : SAT\_2010 data (State SAT score from 2010)**

1. **상기 데이터에 대해 분석을 진행하라.**
   1. **해당 데이터에서 중복 데이터가 존재하는가? 존재한다면 해당 변수명을 명시하라.**



하나의 행이 중복되는 데이터가 있는지 duplicated()함수를 통해 알아본 결과로는 중복 데이터가 존재하지 않는다.

* 1. **해당 데이터에서 문자 변수는 총 몇 개가 있는가?**



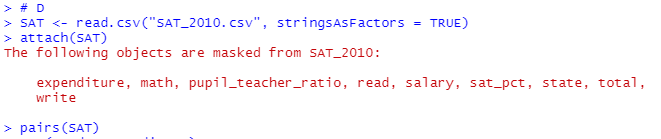
타입이 chr인 문자 변수는 state 1개이다.

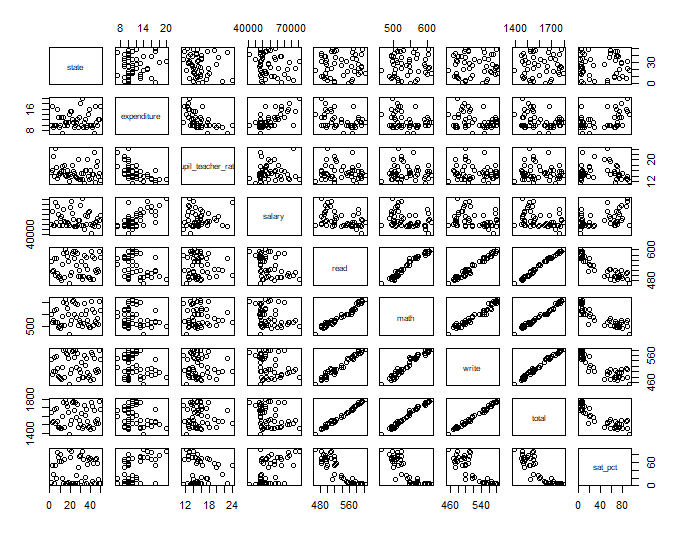
* 1. **결측치가 있는가? 있다면 몇 개가 있는가?**



Is.na()함수를 통해 결측치가 있는지 확인하였고, 이를 sum()함수를 통해 몇 개 인지 출력해본 결과 0개의 결측치가 있다(결측치가 없음).

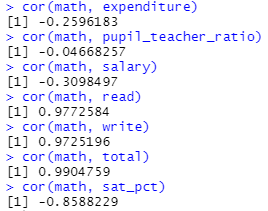
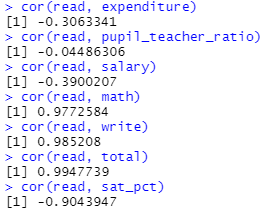
* 1. **변수 read와 변수 math와의 상관계수가 0.9보다 큰 변수를 모두 찾아라.**





직관적으로 보았을 때, read와 math는 write와 total 변수와 상관계수가 커 보인다.

이를 직접 값으로 출력해보면,



read의 경우 상관계수가 0.9보다 큰 변수는 math, write, total이고,

write의 경우 상관계수가 0.9보다 큰 변수는 read, write, total이다.

따라서 read와 write 두 변수 모두에게서 상관계수가 0.9보다 큰 변수는 write와 total이다.

**데이터설명 : happiness\_2020.csv**

country name: 국가명 (153개 data, 19개의 columhn)

regional indicator: 국가가 속한 지역

ladder score: 표본집단이 응답한 행복지수(10에 가까울수록 행복하고, 0에 가까울수록 행복하지 않음을 의미)

logged GDP per capita: 1인당 GDP가 행복지수 계산에 기여하는 정도

Social support: 사회적 지원이 행복지수 계산에 기여하는 정도

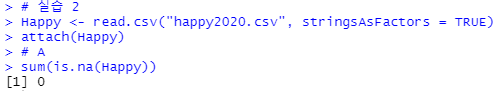
Healthy life expectancy: 기대 수명이 행복지수계산에 기여하는 정도

Freedom to make life choices: 자유의사결정 정도가 행복지수계산에 기여하는 정도

Generosity: 사회적 관용정도가 행복지수계산에 기여하는 정도

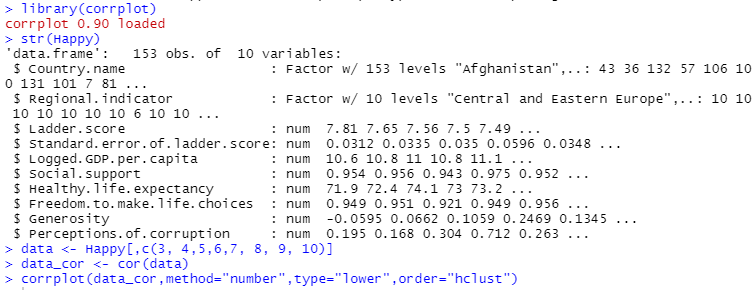
Perceptions of corruption: 부패 인식 지수가 행복지수계산에 기여하는 정도

1. **상기 데이터에 대해 분석을 진행하라.**
   1. **결측치가 있는가? 결측치가 있는지 확인하여라. 있다면 몇 개가 있는가?**

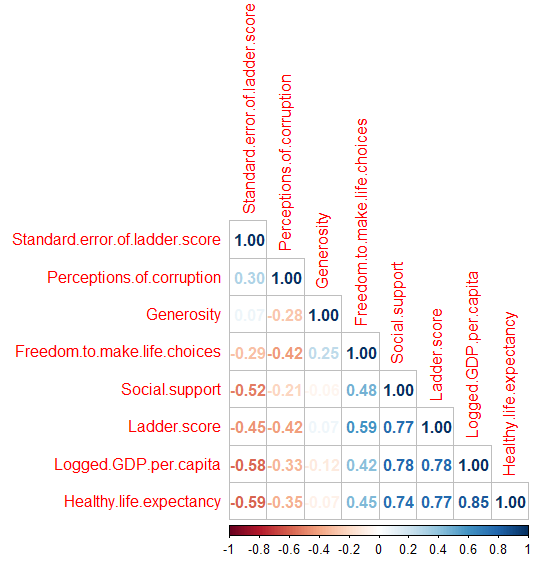


결측치가 없다!

* 1. **행복도(ladder score)에 가장 큰 영향을 미치는 요소 세가지를 상관 분석을 통해서 도출해내라(corrplot 사용 시도 추천). 도출 후, 각각의 산점도와 상관계수를 명시하라.**

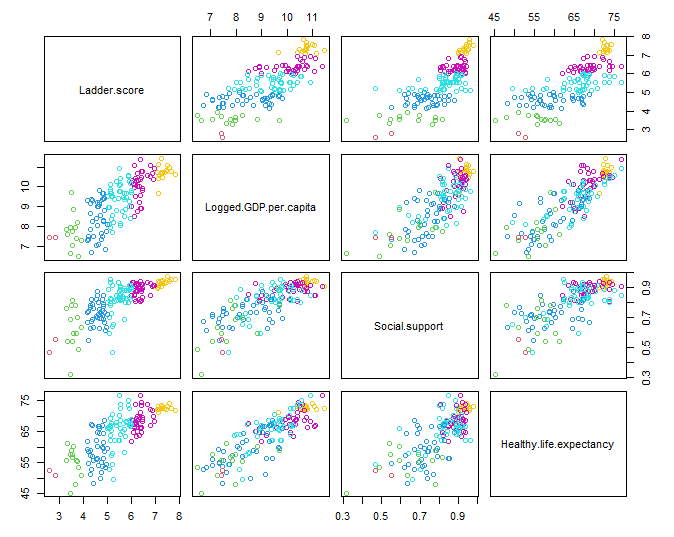


corrplot을 사용하기 위해 패키지를 설치한 후 library를 사용한다. 그래프를 그리기 위해 Factor변수를 제외하고 나머지 num형태의 변수들만 따로 모은 후, cor()함수를 통해 pearson 상관계수를 계산해준다. corrplot을 이용해서 계수가 숫자형태로 나오도록 그래프를 그려준다.



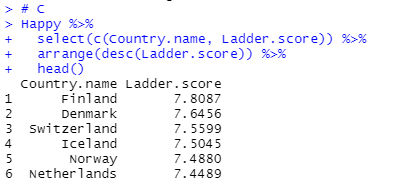
행복도(Ladder.score)에 가장 큰 영향을 미치는 요소는 값이 큰 순서대로 Logged.GDP.per.capita, Support.support, Healthy.life.expectancy 이며 각각의 상관계수 값은 0.78, 0.77, 0.77 이다.



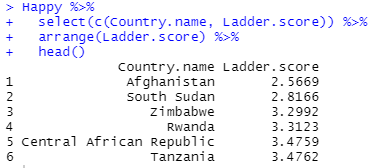


각각의 산점도는 다음과 같다.

* 1. **행복지수가 가장 높은 나라와 낮은 나라는 어디이며 각각의 행복지수는 몇인가?**

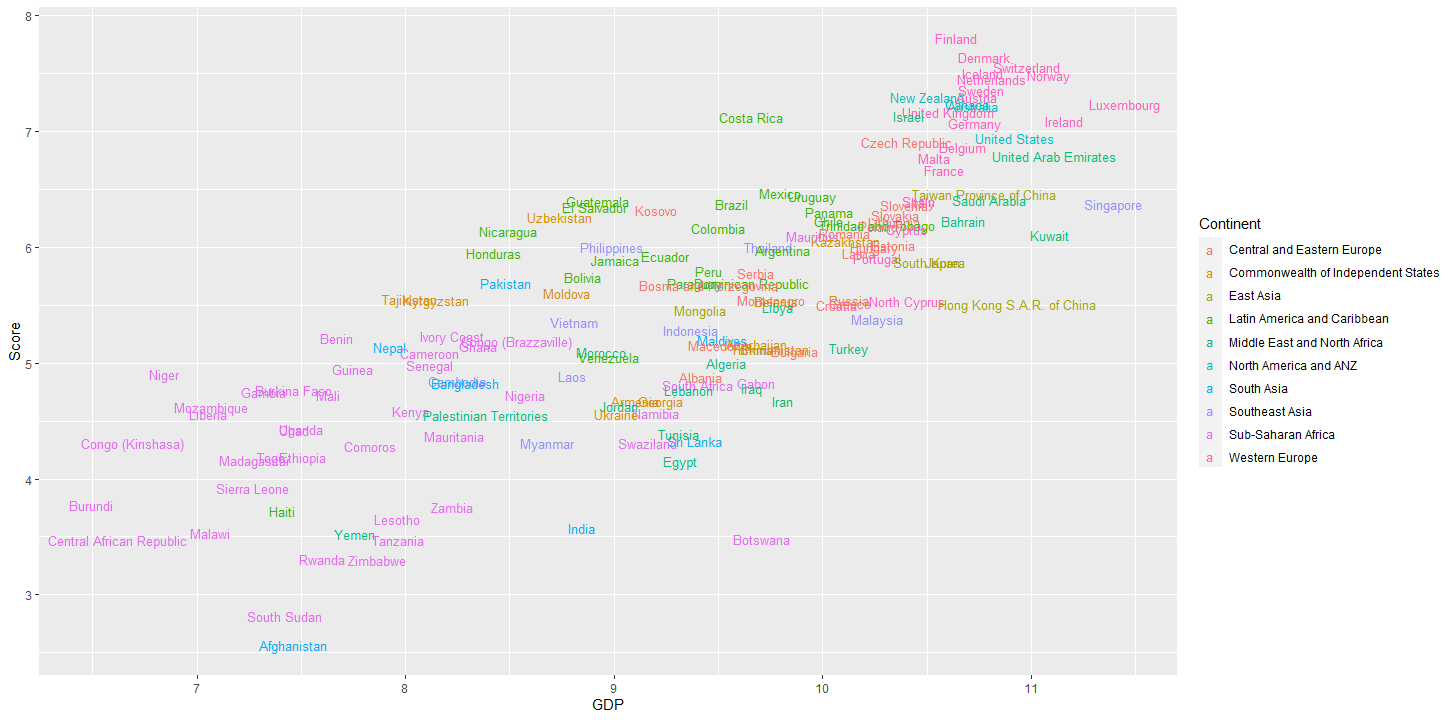


행복지수가 가장 높은 나라는 Finland이고 행복지수는 7.8087이다.



행복지수가 가장 낮은 나라는 Afghanistan이고 행복지수는 2.5669이다.

* 1. **아래의 그래프를 해석하라.**

****

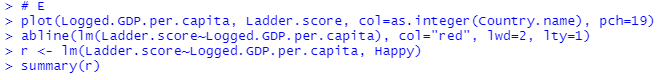
Score라고 기입되어 있는 Y축은 행복지수를 의미하는 것 같고, X축은 GDP를 나타낸다. 그리고 각각의 나라 이름으로 데이터 포인트가 찍혀있으며, 글자 색은 나라를 대륙별로 구분지은 것이다. 이러한 데이터로 볼 때, 전체적으로 상관계수가 0보다는 1에 가까운 형태를 띄고 있으며 이는 GDP가 높은 국가일수록 행복지수가 높음을 알 수 있다. 게다가 Western Europe에 속하는 나라의 경우 대부분이 GDP가 높고 행복지수도 높은 반면, Sub-Saharan Africa에 속하는 나라의 경우 대부분이 GDP가 낮고 행복지수 또한 낮은 것으로 보인다.

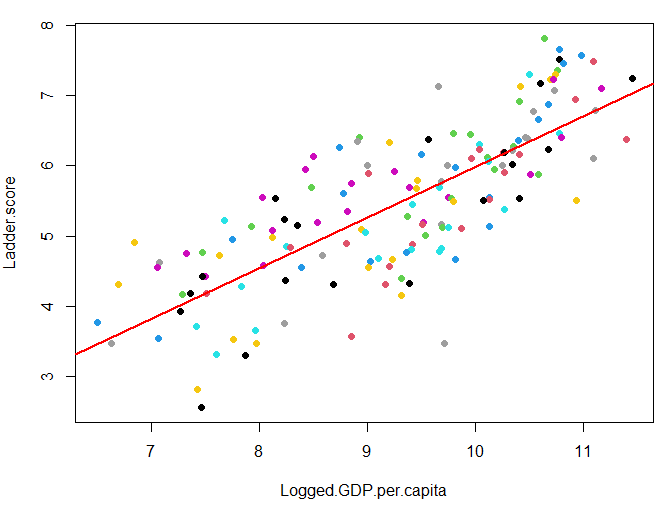
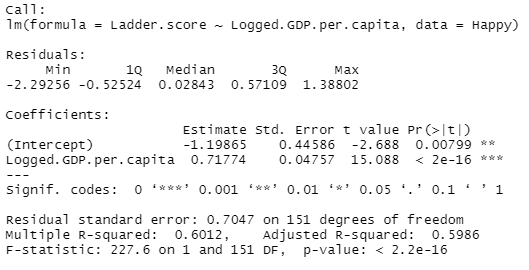
신기한점은 Singapore의 경우 같은 대륙의 나라들에 비해 GDP 수치는 굉장히 높은 것으로 보이는데 반해, 행복지수는 같은 대륙의 주변 나라들보다 조금 더 높은 수준이다.

또한 Luxembourg의 경우가 가장 GDP가 높은데, 그렇다고 행복지수가 가장 높지는 않다.

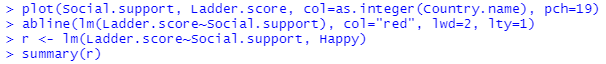
그리고 Afganistan의 경우 동일한 대륙의 다른 나라들에 비해 GDP 수치와 행복 지수 모두 월등히 낮다.

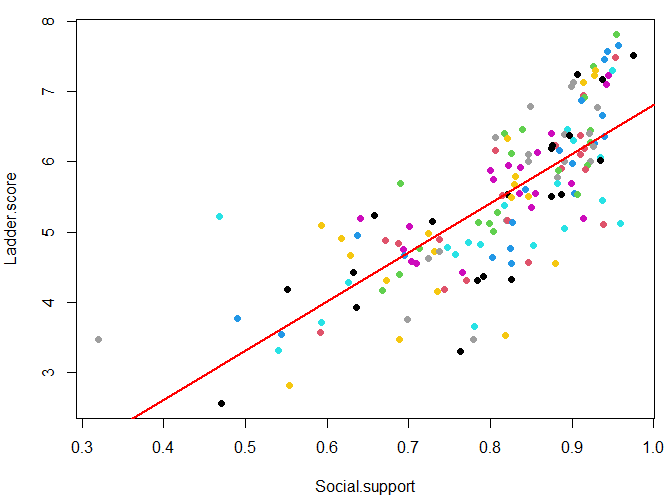
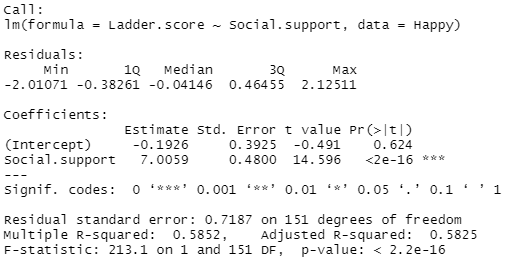
* 1. **B에서 구한 세가지 요소를 이용하여 종속변수를 행복도로 두고 회귀모형을 수행하고 해석하라**



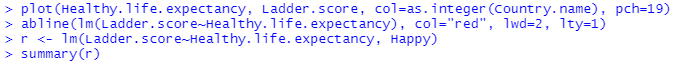
 

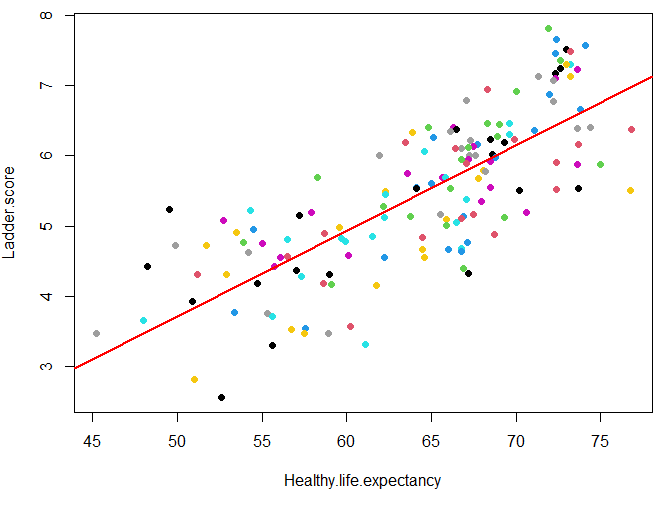
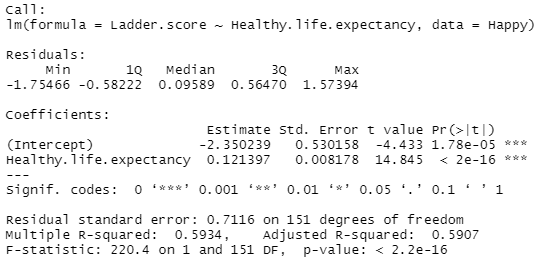
선형회귀식은 y(행복지수) = -1.19865 + 0.71774(1인당 GDP) 이고, R^2 값은 0.6012이다. 결정계수가 0.6정도 이므로 1에 많이 가깝지는 않지만 상대적으로 독립변수들에 의해 적합한 부분이 어느정도 있다고 볼 수 있다.



선형회귀식은 y(행복지수) = -0.1926 + 0.70059(사회적 지원) 이고, R^2 값은 0.5852이다. 결정계수가 0.58정도 이므로 1에 많이 가깝지는 않지만 상대적으로 독립변수들에 의해 적합한 부분이 어느정도 있다고 볼 수 있다.



선형회귀식은 y(행복지수) = -2.350239 + 0.121397(기대 수명) 이고, R^2 값은 0.5934이다. 결정계수가 0.59정도 이므로 1에 많이 가깝지는 않지만 상대적으로 독립변수들에 의해 적합한 부분이 어느정도 있다고 볼 수 있다.

1. **해당 아래 코드를 참고하여 word cloud를 출력하라. (추가 문제)**

filePath <- "http://www.sthda.com/sthda/RDoc/example-files/martin-luther-king-i-have-a-dream-speech.txt"

text <- readLines(filePath)