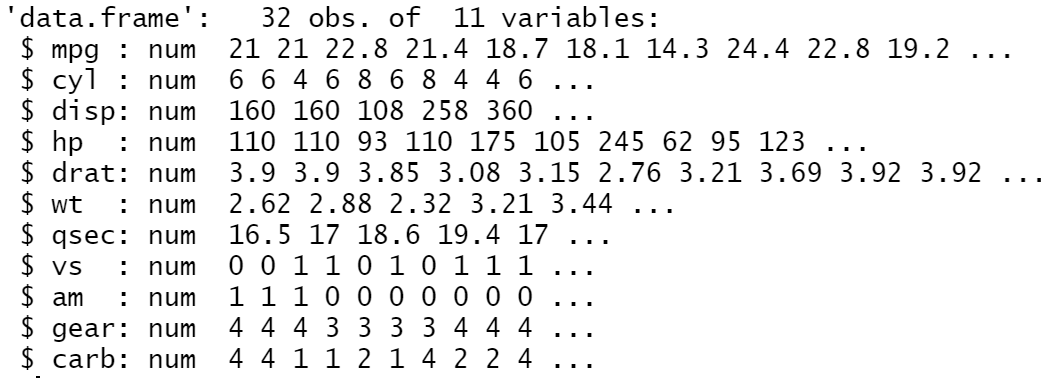
**HW1 (Due date : July.12 PM 19:59, upload in web)**

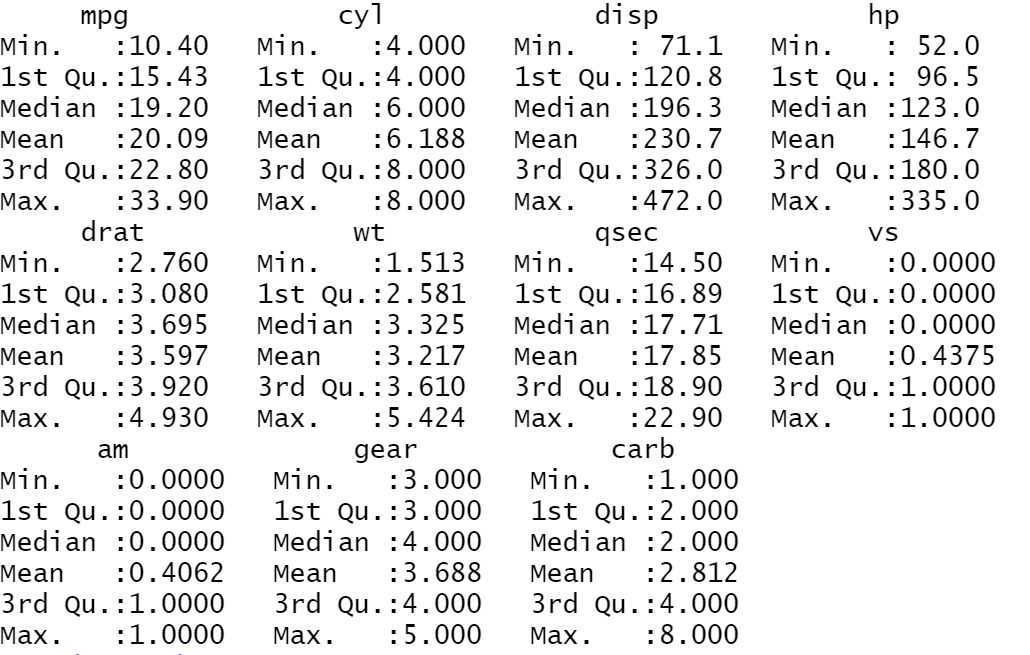
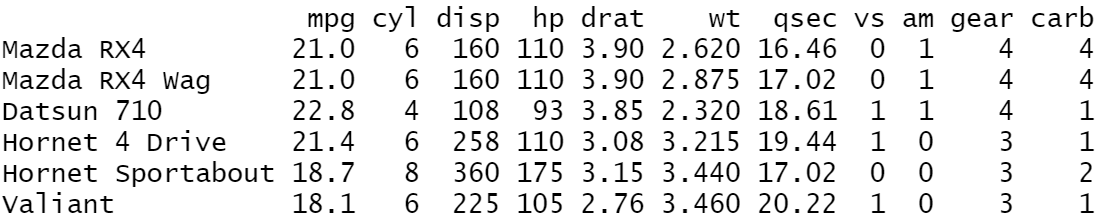
**Submit the homework named as HW1\_이름.doc, HW1\_이름.R (코드도 첨부)**

**문제**

1. **다음은 autompg 데이터에 대한 요약통계치를 나타내는 대표적인 4가지 방법이다. 각각은 dim, head, str, summary 함수의 결과이다. 어떤 함수의 결과인지 순서대로 적어라.**

**(데이터는 UCI Dataset에서 인용)**

****

**  
**

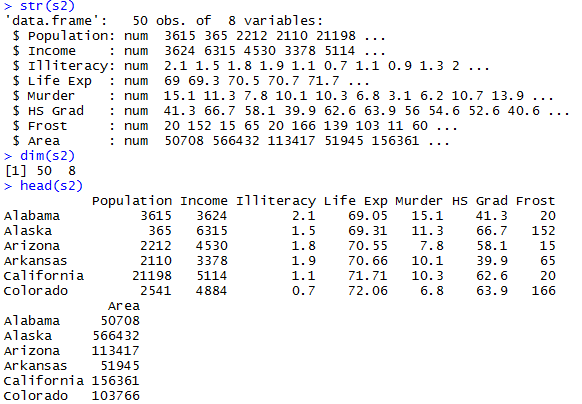
****

1. **데이터 내 그룹별 요약통계치를 구하고 싶을 때 사용하는 함수는?**

**(1) filter (2) select (3) mutate (4) group\_by & summarize**

1. **다음의 결과를 보고 답하여라**

**(데이터는 U.S. Department of Commerce, Bureau of the Census (1977) County and City Data Book에서 인용)**

****

**3-1. s2라는 데이터의 관측치수는 몇 개인가?**

**3-2 s2 데이터에는 factor변수가 있는가?**

**3-3. s2 데이터에는 변수가 몇 개 있는가?**

1. **아래의 데이터에서 성별 평균 키를 뽑는 코드는? (데이터 테이블 명은 data이다.)**

**Height Weight Gender**

**Alice 165 55 Female**

**Bruce 170 65 Male**

**Catherine 155 50 Female**

**David 185 110 Male**

**(1) summarize( group\_by(data, Gender), result=mean(Height))**

**(2) data %>% group\_by(Height) %>% summarize(result =mean(Gender))**

**(3) filter(data,Gender = c("Male, Female)) %>% summarize(result =mean())**

**(4) data %>% select(Height) %>% summarize(result =mean(Gender))**

**실습**

**데이터설명 : The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes.**

* **mpg: continuous**
* **cylinders: multi-valued discrete**
* **displacement: continuous**
* **horsepower: continuous**
* **weight: continuous**
* **acceleration: continuous**
* **model year: multi-valued discrete**
* **origin: multi-valued discrete**
* **car name: string (unique for each instance**

1. **autompg data**

**A. autompg data에서 실린더(cyl)가 8인 데이터만 추출하고 horsepower의 평균을 구하여라. 실린더(cyl)이 4인 데이터를 추출하여 평균을 구하고 비교하라.**

**B. autompg data에서 carname에 ‘chevrolet’이 들어간 차량의 데이터를 추출하라. 몇 개의 데이터가 존재하나? (hint : grepl(carname, ‘chevrolet’)을 사용하면 carname에 ‘chevrolet’이 들어간 차량의 데이터를 알 수 있다.)**

**데이터설명 : Major League Baseball Data from the 1986 and 1987 seasons.**

**A data frame with 322 observations of major league players on the following 20 variables.**

* **AtBat: Number of times at bat in 1986**
* **Hits: Number of hits in 1986**
* **HmRun: Number of home runs in 1986**
* **Runs: Number of runs in 1986**
* **RBI: Number of runs batted in in 1986**
* **Walks: Number of walks in 1986**
* **Years: Number of years in the major leagues**
* **CAtBat: Number of times at bat during his career**
* **CHits: Number of hits during his career**
* **CHmRun: Number of home runs during his career**
* **CRuns: Number of runs during his career**
* **CRBI: Number of runs batted in during his career**
* **CWalks: Number of walks during his career**
* **League: A factor with levels A and N indicating player's league at the end of 1986**
* **Division: A factor with levels E and W indicating player's division at the end of 1986**
* **PutOuts: Number of put outs in 1986**
* **Assists: Number of assists in 1986**
* **Errors: Number of errors in 1986**
* **Salary: 1987 annual salary on opening day in thousands of dollars**
* **NewLeague: A factor with levels A and N indicating player's league at the beginning of 1987**

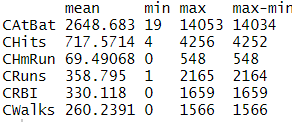
1. **다양한 야구 지표를 계산해본다.**
   1. **각 선수별 타율을 계산하여 기존 테이블에 넣어라**

**타율 =- Hits/AtBat**

* 1. **각 선수별 출루율을 계산하여 기존 테이블에 넣어라.**

**출루율 = (Hits + Walks)/(AtBat + Walks)**

* 1. **타율과 출루율의 평균, 중앙값을 계산하라.**
  2. **출루율과 평균이 모두 평균보다 높거나 같은 선수는 전체에서 몇 퍼센트의 비율을 차지하는가?**
  3. **기존 Data에서 CAtBat CHits CHmRun CRuns CRBI CWalks만 선택하여 새로운 데이터를 만들어라.**
  4. **E의 결과에 대해 mean, min, max와 min과 max의 차이를 보여주는 표를 만들어라.**

****