**R 프로그래밍 과제**

**22100748 최윤영**

mpg data 관련해서 참고

|  |  |
| --- | --- |
| **mpg data** | **설명** |
| displ | 배기량 |
| hwy | 고속도로 연비 |
| cty | 도시 연비 |
| manufacturer | 제조 회사 |
| class | 차종 |

1. **Data preprocessing**

**문제 1**

**1.1** 배기량(displ)에 따라 고속도로 연비 (hwy)가 다른지 알아 보고자 한다. 배기량 (displ) 이 4 이하인 자동차와 5이상인 자동차 중 어떤 자동차의 고속도로 연비가 평균적으로 더 높은지 각 평균값을 출력하여라.

|  |
| --- |
|  |

**1.2** 어떤 회사 (manufacturer) 자동차의 (hwy)가 가장 높은가 알고자 한다. hwy 평균 이 가장 높은 회사 3곳을 출력하여라.

|  |
| --- |
|  |

1. **Using ggplot2 package**

**문제 2번**

**2.1**  mpg 데이터에서 도시 연비 (cty) 고속도로연비 (hwy) 간의 관계 를 파악하고자 한다. x축에 도시 연비, y축에 고속도로 연비로 된 산점도(scatter plot graph)를 그려라.

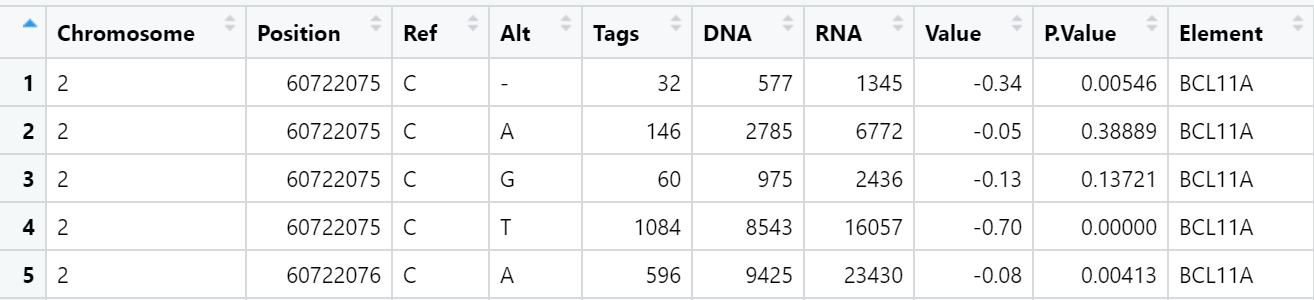
|  |
| --- |
|  |

1. **MPRA 데이터 분석하기**

**문제 3번**

MPRA 데이터에서 position은 염기 위치, Ref는 기존의 서열, Alt는 돌연변이로 인해 치환된 염기(‘-’는 deletion을 의미), Value는 돌연변이로 인해서 기존 대비 변하게 된 발현량의 정도를 뜻한다. MPRA 데이터에서 돌연변이의 종류에 따른 발현량의 변화 양상을 알아보고자 한다.

GRCh27\_ALL.csv 데이터 참고 (자세한 내용은 RStudio에서 View()를 통해 확인)



|  |
| --- |
| [Tip] 변수 추가시에 ifelse문을 사용하기를 권장한다.  ex 1) age가 23이상이면 old 아니면 young인 test 변수를 DF에 추가  DF <- DF %>% mutate(test=ifelse(DF$age>=23,’old,’young))  ex 2) 중첩조건문-score가 90이상이면 A, 80이상은 B, 외에는 C를 주는 grade변수를 DF에 추가  DF <- DF %>% mutate(grade=ifelse(DF$score >=90, ‘A’, iflese(DF$score >=80, ‘B’, ‘C’ ))) |

* 1. GRCh27\_ALL.csv 데이터를 읽어서 MPRA\_data에 저장하기 (read.csv() 사용).
  2. Ref가 pyrimidine과 purine 중 어떤 타입인지를 뜻하는 Ref\_type 변수 생성 (위 노란색 상자에 있는 Tip 내용 참고).
  3. Alt가 pyrimidine, purine, deletion 중 어떤 타입인지를 뜻하는 Alt\_type 변수 생성. (Tip 내용 참고)

‘-’는 deletion을 의미한다.

* 1. transition, transversion, deletion 중 어떤 돌연변이인지를 뜻하는 mutation\_type 변수 생성 (Tip 내용 참고)
     + 기존의 서열과 돌연변이로 치환된 염기가 같다면 transition (hint: step b와 c에서 추가한 변수를 이용할 것)
     + Alt 값이 deletion일 경우는 deletion
     + 그 외의 경우는 transversion
  2. Value가 0보다 크거나 같으면 increase, 작으면 decrease인 effect 변수 생성
  3. ggplot을 이용하여 mutation\_type에 따른 Value의 평균값을 나타내는 그래프를 그리기(평균막대그래프, geom\_col() 이용).

|  |
| --- |
|  |

* 1. ggplot을 이용하여 increase와 decrease 중에서 어떤 경우의 빈도수가 더 많은지를 나타내는 그래프를 그리기 (빈도막대그래프, geom\_bar() 이용).

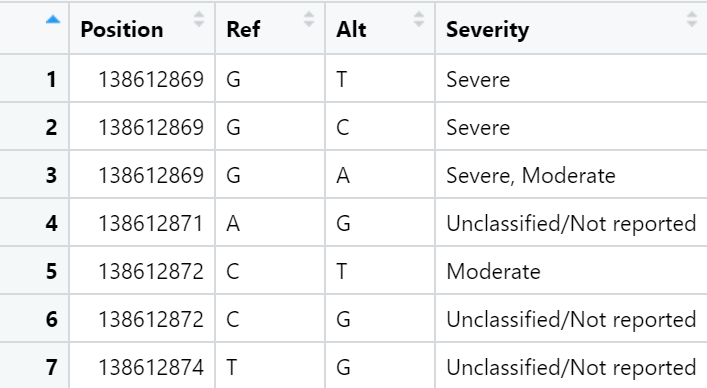
|  |
| --- |
|  |

1. **Hemophilia 환자 데이터와 함께 분석하기**

**문제 4번** (총 15점, 각각 5점씩)

F8, F9 database - F9.csv에는 실제 혈우병 환자의 F9 promoter에서의 돌연변이의 정보와 병의 중증도가 포함되어 있다.

F8, F9 database - F9.csv 데이터 참고 (자세한 내용은 RStudio에서 View()를 통해 확인)



1. F8, F9 database - F9.csv 데이터를 읽어서 F9\_medical\_data에 저장 (read.csv 사용)
2. F9\_data<-merge(F9\_medical\_data,MPRA\_data,by=c("Position","Ref","Alt"))를 사용하여 MPRA데이터와 Hemophilia데이터를 합친 F9\_data를 생성하기.
3. ggplot을 이용하여 severity에 따른 Value의 평균값을 나타내는 그래프를 그리기 (geom\_col() 이용)

|  |
| --- |
|  |