인문학 텍스트 마이닝

NA(Missing Value) Handling

NA Handling

- NA(Not avaliable)
  - 값이 누락되거나 값이 없는 값을 나타내는 문자

- 예제1
  - 변수 생성

```
> X<-c(1,2,3,4,5,6,7,8,NA)
> X
[1]  1  2  3  4  5  6  7  8 NA
```

  - NA값으로 변환

```
> X[X==2]<-NA
> X
[1]  1 NA  3  4  5  6  7  8 NA
```

  - 변수 요약

```
> summary(X)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1.00    2.75    4.50    4.50    6.25    8.00       1
```

  - 변수 연산하기

```
> sum(X)
[1] NA
> mean(X)
[1] NA
> sum(X,na.rm=T)
[1] 34
> mean(X,na.rm=T)
[1] 4.857143
```

NA Handling

- 예제2
- 남녀간의 영어,수학 점수를 나타내는 데이터세트 생성

```
> Eng<-c(34,45,56,67,78,89,NA)
> Math<-c(98,NA,87,76,65,54,43)
> Gender<-c("M","F","M","F","M","M","M")
> Test<-data.frame(Eng=Eng,Math=Math,Gender,Gender)
> Test
  Eng Math Gender Gender.1
1  34   98      M        M
2  45   NA      F        F
3  56   87      M        M
4  67   76      F        F
5  78   65      M        M
6  89   54      M        M
7  NA   43      M        M
```

- 데이터 확인

```
> str(Test)
'data.frame':   7 obs. of  4 variables:
 $ Eng     : num  34 45 56 67 78 89 NA
 $ Math    : num  98 NA 87 76 65 54 43
 $ Gender  : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 2
 $ Gender.1: Factor w/ 2 levels "F","M": 2 1 2 1 2 2 2
```

NA Handling

- NA를 포함한 행을 제거한 데이터 세트 생성

```
> na.omit(Test)
  Eng Math Gender Gender.1
1  34   98      M        M
3  56   87      M        M
4  67   76      F        F
5  78   65      M        M
6  89   54      M        M
```

- Test 데이터 요약

```
> summary(Test)
      Eng              Math       Gender
 Min.   :34.00   Min.   :43.00   F:2
 1st Qu.:47.75   1st Qu.:56.75   M:5
 Median :61.50   Median :70.50
 Mean   :61.50   Mean   :70.50
 3rd Qu.:75.25   3rd Qu.:84.25
 Max.   :89.00   Max.   :98.00
 NA's   :1       NA's   :1
```

- 평균치 삽입법을 활용한 NA데이터 조작

```
> install.packages("gam")
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/gam_1.09.1.tgz'을 시도합니다
Content type 'application/x-gzip' length 304040 bytes (296 Kb)
URL을 열었습니다
==================================================
downloaded 296 Kb


The downloaded binary packages are in
        /var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//RtmpV2osnU/downloaded_packages
> library(gam)
> na.gam.replace (Test)
   Eng Math Gender
1 34.0 98.0      M
2 45.0 70.5      F
3 56.0 87.0      M
4 67.0 76.0      F
5 78.0 65.0      M
6 89.0 54.0      M
7 61.5 43.0      M
```

NA Handling

- 영어와 수학 점수만으로 구성된 데이터 세트 생성

```
> Test2<-Test[,c("Eng","Math")]
> Test2
  Eng Math
1  34   98
2  45   NA
3  56   87
4  67   76
5  78   65
6  89   54
7  NA   43
```

- 데이터 세트 연산하기

```
> apply(Test2,2,mean)
 Eng Math
  NA   NA
> apply(Test2,2,mean,na.rm=TRUE)
 Eng Math
61.5 70.5
```

Outliers Handling

Outliers Handling

- Outliers
  - 데이터 안에 존재하는 이상치로 데이터의 성질에 큰 영향을 미친다.

- 예제1
  - 라이브러리 설치

```
> install.packages("outliers")
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/outliers_0.14.tgz'을 시도합니다
Content type 'application/x-gzip' length 50370 bytes (49 Kb)
URL을 열었습니다
==================================================
downloaded 49 Kb


The downloaded binary packages are in
        /var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//Rtmp8qGMiY/downloaded_packages
> library(outliers)
```

  - 수치형 변수 추출

```
> Test_1=Test[,c(1,2)]
>
```

Outliers Handling

- 표준화 관련 수식(Formula)

$$Z_i = \frac{W_i - \bar{W}}{S_W} \qquad \bar{W} = \frac{\sum_{i=1}^{n} W_i}{n} \qquad S_W = \sqrt{\frac{\sum (W_i - \bar{W})^2}{n}}$$

- NA값에 평균치 삽입

```
> library(gam)
필요한 패키지를 로딩중입니다: splines
Loaded gam 1.09.1

>
> na.gam.replace (Test_1)
   Eng Math
1 34.0 98.0
2 45.0 70.5
3 56.0 87.0
4 67.0 76.0
5 78.0 65.0
6 89.0 54.0
7 61.5 43.0
```

- 데이터 표준화 시키기

```
> X <- scores(na.gam.replace (Test_1), type=c("z"))
>
> b=scale(na.gam.replace (Test_1))
>
> as.data.frame(b)
         Eng       Math
1 -1.4638501  1.4638501
2 -0.8783101  0.0000000
3 -0.2927700  0.8783101
4  0.2927700  0.2927700
5  0.8783101 -0.2927700
6  1.4638501 -0.8783101
7  0.0000000 -1.4638501
```

Outliers Handling

- 데이터 필터링 하기

```
> install.packages("dplyr")
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/dplyr_0.4.1.tgz'을 시도합니다
Content type 'application/x-gzip' length 3781115 bytes (3.6 Mb)
URL을 열었습니다
==================================================
downloaded 3.6 Mb


The downloaded binary packages are in
        /var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//Rtmp8qGMiY/downloaded_packages
>
> library(dplyr)

다음의 패키지를 부착합니다: 'dplyr'

The following object is masked from 'package:stats':

    filter

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

>
> filter(X, Eng <= 1, Math <= 1)
         Eng       Math
1 -0.8783101  0.0000000
2 -0.2927700  0.8783101
3  0.2927700  0.2927700
4  0.8783101 -0.2927700
5  0.0000000 -1.4638501
```

NA Handling

- 실습 문제
- 다음은 어느 한 반의 기말고사 성적이다.

	A	B	C	D	E	F	G	H
1	Num	Name	Eng	Math	Korean	Society	Science	Gender
2	1	Amy	97	74	93	96	76	F
3	2	Alexis	75	90	80	84	88	M
4	3	Lexi	86	76	88	90	71	F
5	4	Katie	88	77	85	89	74	F
6	5	Ivy	87	84	90	88	83	F
7	6	Teddy	83	97	75	86	89	M
8	7	Rocky	82	93	79	83	91	M
9	8	Becca	89	73	95	83	80	F
10	9	Taylor	80	94	73	81	94	M
11	10	Sam	79	95	72	78	93	M
12	11	Lauren	90	77	88	86	87	F
13	12	Iris	93	79	86	87	77	F
14	13	Simon	84	89	78	77	90	M
15	14	Tom	83	94	77	74	89	M
16	15	Becca	96	69	89	90	78	F
17	16	Cece	94	74	95	93	74	F
18	17	Kabin	77	95	74	85	95	M
19	18	Robert	86	90	78	82	94	M
20	19	Rebecca	96	79	90	86	67	F
21	20	Garin	85	95	83	80	90	M

NA Handling

- 실습 문제

- (1) 각 과목의 총합을 구하고, 분석에 사용한 R 코드를 적으시오.

- (2) 각 과목의 평균을 구하고, 분석에 사용한 R 코드를 적으시오.

- (3) 각 과목별로 성적이 가장 높은 학생은 누구인지 찾고, 분석에 사용한 R 코드를 적으시오. (결과 값에 각 과목의 성적과 학생의 이름이 모두 출력되게 하시오.)