Sankey Diagram

Sankey 다이어그램은 두 개 혹은 두 개 이상의 변수간의 데이터 흐름을 잘 보여주는 그래프이다. 각각의 변수 항목들은 네모 박스로 표현하고 데이터가 연관된 항목간의 데이터 양에 따라 굵거나 다른 선으로 이어지는 형태로 표현되는 그래프로 비교적 최근부터 사용되기 시작한 그래프 형태이다.

아쉽게도 R에서 그래프를 그리는데 가장 많이 사용되는 `ggplot2` 는 아직까지 Sankey Diagram을 지원하지 못한다. 따라서 Sankey 다이어그램을 생성하기 위해서는 `plotly` 나 `networkD3` 패키지를 사용할 수 있다.

`plotly` 와 `networkD3` 패키지로 작성한 Sankey 다이어그램은 모두 대화형(interactive) 그래프로 Sankey 다이어그램이 생성되기 때문에 별상에서는 마우스 포인터의 위치에 따라 해당 내용이 하면에 표기된다.

Sankey 다이어그램을 생성하기 위해서는 일단 Sankey 다이어그램을 생성하기 위한 데이터를 세팅할 필요가 있다.

이번 포스트에서는 교육통계 서비스 홈페이지 https://kess.kedi.re.kr (https://kess.kedi.re.kr)에서 제공하는 직업계고 시도별 유형별 취업현황 데이터 (https://kess.kedi.re.kr/post/67247177?code=&words=&since=&until=&page=0&itemCode=04&menuId=m_02_04_03_01)를 사용하였다.

데이터 로딩 및 전처리

직업계고 시도별 유형별 취업현황 데이터 (https://kess.kedi.re.kr/post/67247177?code=&words=&since=&until=&page=0&itemCode=04&menuId=m_02_04_03_01)에서 다운받은 파일(직업계고) 시도유형별 취업현황_2020.xlsx의 데이터를 로딩하면 다음과 같다.

```
library(readxl)

df <- read_excel('./(직업계고) 시도_유형별 취업현황_2020.xlsx', skip = 2, na = '-', sheet = 1, col_types = c("text",
  'text', rep('numeric', 10)), col_names = F)

colnames(df) <- c('지역', '종류', '졸업자_계', '졸업자_남', '졸업자_여', '취업자_계', '취업자_남', '취업자_여', '진학자_계
, '진학자_남', '진학자_여', '입대자_계', '입대자_남', '입대자_여', '제외인정자_계', '제외인정자_남', '제외인정자_여', '미취업
자_계', '미취업자_남', '미취업자_여')

head(df)
```


```
## # A tibble: 6 x 20
##   지역   종류   졸업자_계 졸업자_남 졸업자_여 취업자_계 취업자_남 취업자_여
##   <chr>  <chr>      <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 01_서울 마이스터고   584       411       173       418       279       139
## 2 02_부산 마이스터고   549       539        10       279       273         6
## 3 03_대구 마이스터고   563       525        38       290       272        18
## 4 04_인천 마이스터고   265       224        41       117        93        24
## 5 05_광주 마이스터고   157       122        35       103        81        22
## 6 06_대전 마이스터고   264       247        17       178       166        12
## # ... with 12 more variables: 진학자_계 <dbl>, 진학자_남 <dbl>,
## #   진학자_여 <dbl>, 입대자_계 <dbl>, 입대자_남 <dbl>, 입대자_여 <dbl>,
## #   제외인정자_계 <dbl>, 제외인정자_남 <dbl>, 제외인정자_여 <dbl>,
## #   미취업자_계 <dbl>, 미취업자_남 <dbl>, 미취업자_여 <dbl>
```


Sankey 다이어그램을 생성하기 위해서는 세가지 데이터가 필요하다.

첫번째는 (네모 박스로 표현되는) 각각 노드의 이름,

두번째는 각각의 노드들이 연결되는 링크에 대한 정보,

세번째는 링크의 굵기가 표현할 데이터 정보이다.

이를 적절히 추출하기 위해 앞서 모일한 데이터를 다음과 같이 변환하였다.

```
library(tidyverse)
sankey <- df %>%
  ## '지역'열과 '종류'열 중에 na가 아닌 행만 선택한다.
  filter(!is.na(지역) == FALSE, !is.na(종류) == FALSE) %>%
  ## 열 중에서 '지역','종류','졸업자','취업자 '남', '여'로 끝나는 열을 제외
  select(-c(지역, 종류), 졸업자, ends_with('남'), ends_with('여')) %>%
  ## 종류 별로 사용하여 group
  group_by(종류) %>%
  ## 전체 갯수 대신 sum 을 적용
  summarise_all(sum) %>%
  ## 데이터를 가로의 변수 -> 세로의 변수로 변경
  gather(구분, '학생수', -1) %>%
  ## 종류 열과 구분 열을 factor로 변환
  mutate(종류 = fct_relevel(종류, '마이스터고', '특성화고', '일반고_직업반'),
         구분 = fct_relevel(구분, '취업자', '진학자', '입대자', '제외인정자', '미취업자')) %>%
  ## 종류 별과 구분 별로 정렬
  arrange(종류, 구분)

head(sankey)
```


```
## # A tibble: 6 x 3
##   종류       구분      학생수
##   <fct>      <fct>      <dbl>
## 1 마이스터고  취업자      3510
## 2 마이스터고  진학자       297
## 3 마이스터고  입대자       394
## 4 마이스터고  제외인정자     48
## 5 마이스터고  미취업자     1417
## 6 특성화고    취업자     20785
```


위의 데이터셋서 앞서 설명한 세가지 데이터를 뽑아내겠다.

1. 노드의 이름

노드의 이름은 좌측 노드(from)와 우측 노드(to)의 이름을 벡터로 만들었다. 좌측 노드는 종류 열, 우측 노드는 구분 열로 정리했기 때문에 각각의 열을 벡터로 변환하면 쉽게 얻을 수 있다.

```
from <- unique(as.character(sankey$종류))
from
```

```
## [1] "마이스터고"   "특성화고"   "일반고_직업반"
```

```
to <- unique(as.character(sankey$구분))
to
```

```
## [1] "취업자"     "진학자"     "입대자"     "제외인정자" "미취업자"
```

```
c(from, to)
```

```
## [1] "마이스터고"   "특성화고"   "일반고_직업반" "취업자"
## [5] "진학자"       "입대자"     "제외인정자"   "미취업자"
```

2. 노드 링크 정보

노드 링크 정보는 앞서 생성한 노드의 이름 벡터의 인덱스 정보를 사용해서 **source와 target**을 설정한다. 여기서 중요한 것이 R에서 사용하는 인덱스는 일반적으로 1부터 시작하지만 여기서 사용하는 인덱스는 0부터 시작한다는 것이다. 앞서 생성한 벡터에서 가장 앞에 위치한 '특성화고'는 인덱스 0이므, 두번째인 '마이스터고'는 인덱스 1로 사용한다. 여기서 그리고자 하는 Sankey 다이어그램은 특성화고(0)-취업자(3), 특성화고(0)-진학자(4), 특성화고(0)-입대자(5), 특성화고(0)-제외인정자(6), 특성화고(0)-미취업자(7), 마이스터고(1)-취업자(3), ... , 일반고_직업반(2)-미취업자(7) 등의 링크가 필요하다. 따라서 다음과 같이 **source와 target**을 설정할 수 있다.

```
## source
c(rep(0, 5), rep(1, 5), rep(2, 5))
```

```
##  [1] 0 0 0 0 0 1 1 1 1 1 2 2 2 2 2
```

```
## target
c(rep(3:7, 3))
```

```
##  [1] 3 4 5 6 7 3 4 5 6 7 3 4 5 6 7
```

3. 노드 데이터 정보

앞서 설정한 노드 링크 정보는 총 15개이다. 따라서 각 노드 링크마다 할당되어야 하는 데이터의 개수도 15개가 필요하다. 앞서 데이터를 선처리 할 때 노드링크의 순서에 따라 데이터를 정리해 놓았기 때문에 **sankey$학생수**를 바로 사용할 수 있다.

```
## 노드 데이터 정보
sankey$학생수
```

```
##  [1]  3510   297   394    48  1417 20785 35195  1176   864 21483   643  2723
## [13]    15    54  1390
```

위의 데이터들을 정리하자면 노드의 이름과 연계 설정하고 노드의 링크, 노드의 데이터를 설정하는데 노드 0은 '특성화고', 노드 3은 '취업자', 노드 4는 '진학자'이고 **0->3**으로 가는 링크는 **3510**, **0->4**로 가는 링크는 **297**임과 같이 표현된다.

networkD3 를 활용한 sankey diagram

`networkD3` 를 사용해서 Sankey 다이어그램을 생성하는 것은 `plotly` 를 사용할 때와 거의 유사하다. 다만 `plotly` 를 사용해서 Sankey 다이어그램을 생성할 때는 Sankey 다이어그램을 생성하기 위한 세가지 데이터들을 벡터로 사용하였지만 `networkD3` 를 사용할 때는 데이터 프레임의 형태로 저장하여 사용해야 한다는 점이 다르다.

```
library(networkD3)
## 노드와 링크의 저장할 데이터프레임 생성
nodes <- data.frame(
  name = c(from, to)
)

## 노드 링크 정보와 노드 데이터 정보가 저장된 데이터 프레임 생성
links <- data.frame(
  source = c(rep(0, 5), rep(1, 5), rep(2, 5)),
  target = c(rep(3:7, 3)),
  value = sankey$학생수
)

## sankey diagram 생성
sankeyNetwork(Links = links, Nodes = nodes,
              Source = "source", Target = "target",
              Value = "value", NodeID = "name",
              sinksRight=FALSE, fontSize = 12, nodeWidth = 20)
```