HW1-1_201823869_조성우

조성우

2020 4 30

데이터 프레임 전처리

• 첨부된 data 파일 또는 개별적으로 다운로드한 데이터파일 중 1를 선택 • 사용할 데이터를 읽어와서 dataframe 생성 • 변수들의 속성을 데이터에 알맞게 변환하기(문자, 팩터, 숫자 등) • 변수를 파악하는 함수, 계산함수 등 각종 함수 사용하기(통계함수, 서브세팅, 비교연산) • 결측 값이 있는 경우 알맞은 값으로 처리(평균처리, 0으로 처리) • 여러 변수를 활용해 다양한 계산을 통해 dataframe 안에 새로운 변수 생성

```
#-------------------------------------------------
```

지도 시각화

```
library(dplyr) # 파이프함수를 사용하기 위해 dplyr 패키지를 불러옴
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#install.packages("leaflet") #지도 제작을 위해 leaflet 패키지를 불러옴
library(leaflet)
```

```
## Warning: package 'leaflet' was built under R version 3.6.3
```

```
covid_case <- read.csv("http://bitly.kr/C5ykr25ql", stringsAsFactors = F)  # 문자열을 팩터로 변환하지 않는다는 조건으로 해당 링크에서 다루고자 하는 데이터프레임을 가져옵니다.
str(covid_case)
```

```
## 'data.frame':    102 obs. of  8 variables:
##  $ case_id       : int  1000001 1000002 1000003 1000004 1000005 1000006 1000007 1000008 1000
009 1000010 ...
##  $ province      : chr  "Seoul" "Seoul" "Seoul" "Seoul" ...
##  $ city          : chr  "Guro-gu" "Dongdaemun-gu" "Guro-gu" "Eunpyeong-gu" ...
##  $ group         : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ infection_case: chr  "Guro-gu Call Center" "Dongan Church" "Manmin Central Church" "Eunpy
eong St. Mary's Hospital" ...
##  $ confirmed     : int  96 20 20 14 13 10 7 6 109 32 ...
##  $ latitude      : chr  "37.508163" "37.592888" "37.481059" "37.63369" ...
##  $ longitude     : chr  "126.884387" "127.056766" "126.894343" "126.9165" ...
```

```
#전처리----------
covid_case$longitude <- as.numeric(covid_case$longitude)
```

```
## Warning: 강제형변환에 의해 생성된 NA 입니다
```

```
covid_case$latitude <- as.numeric(covid_case$latitude)    ##str(covid_case)로 해당 데이터프레임의 구조를 확인해 본 결과 latitude(위도)/longitude(경도) 가 character로 설정되어 있어 차후 지도에 발생지를 표시하고자 할 때 문제가 될것이므로 ,밑의 as.numeric 과정으로 숫자형 데이터로 변환 시켜줍니다.
```

```
## Warning: 강제형변환에 의해 생성된 NA 입니다
```

```
str(covid_case) #위도,경도가 numeric data로 변환된 것을 볼 수 있습니다.
```

```
## 'data.frame':    102 obs. of  8 variables:
##  $ case_id       : int  1000001 1000002 1000003 1000004 1000005 1000006 1000007 1000008 1000
009 1000010 ...
##  $ province      : chr  "Seoul" "Seoul" "Seoul" "Seoul" ...
##  $ city          : chr  "Guro-gu" "Dongdaemun-gu" "Guro-gu" "Eunpyeong-gu" ...
##  $ group         : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ infection_case: chr  "Guro-gu Call Center" "Dongan Church" "Manmin Central Church" "Eunpy
eong St. Mary's Hospital" ...
##  $ confirmed     : int  96 20 20 14 13 10 7 6 109 32 ...
##  $ latitude      : num  37.5 37.6 37.5 37.6 37.6 ...
##  $ longitude     : num  127 127 127 127 127 ...
```

```
names(covid_case)[c(5,7,8)] <- c("name","lat","long") #편의를 위해 발생지 , 위도 , 경도의 이름을 바꿔줍니다
str(covid_case)
```

```
## 'data.frame':    102 obs. of  8 variables:
##  $ case_id  : int  1000001 1000002 1000003 1000004 1000005 1000006 1000007 1000008 1000009 1
000010 ...
##  $ province : chr  "Seoul" "Seoul" "Seoul" "Seoul" ...
##  $ city     : chr  "Guro-gu" "Dongdaemun-gu" "Guro-gu" "Eunpyeong-gu" ...
##  $ group    : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ name     : chr  "Guro-gu Call Center" "Dongan Church" "Manmin Central Church" "Eunpyeong
St. Mary's Hospital" ...
##  $ confirmed: int  96 20 20 14 13 10 7 6 109 32 ...
##  $ lat      : num  37.5 37.6 37.5 37.6 37.6 ...
##  $ long     : num  127 127 127 127 127 ...
```

```
covid_case
```

case_id <int>	province <chr>	city <chr>	gr… <lgl>	name <chr>
1000001	Seoul	Guro-gu	TRUE	Guro-gu Call Center
1000002	Seoul	Dongdaemun-gu	TRUE	Dongan Church
1000003	Seoul	Guro-gu	TRUE	Manmin Central Church
1000004	Seoul	Eunpyeong-gu	TRUE	Eunpyeong St. Mary's Hospital
1000005	Seoul	Seongdong-gu	TRUE	Seongdong-gu APT
1000006	Seoul	Jongno-gu	TRUE	Jongno Community Center
1000007	Seoul	Jung-gu	TRUE	Jung-gu Fashion Company
1000008	Seoul	from other city	TRUE	Shincheonji Church
1000009	Seoul	-	FALSE	overseas inflow
1000010	Seoul	-	FALSE	contact with patient

1-10 of 102 rows | 1-6 of 8 columns Previous **1** 2 3 4 5 6 … 11 Next

#해당 데이터를 살펴볼 때 위,경도가 NA값으로 들어간 데이터들이 존재하는데, 몇개의 예외를 빼고 대부분의 경우 infection_case를 살펴 볼 때 위/경도가 존재하는 다른데이터들과 달리 특정 발생지를 말하는 것이 아닌 일반화 된 케이스를 말하고 있기 때문에 NA값으로 위치가 특정되지 않은것임을 알 수 있습니다.
#위도/경도에 city값을 기준으로 평균을 내서 작은 등차값을 주어 NA에 할당 할 수도 있겠으나, 해당 마킹 지도의 목적이 사용자에게 코로나 발생지를 알림으로써 주의를 주기위함임이라고 고려할 때 앞에 언급한 전처리는 불필요한데 더해 해석에 혼란을 야기할 수 있다고 생각하여 하지 않겠습니다. 더불어 위도/경도를 0으로 처리한다면 Mark가 지역의 도메인인 대한민국을 벗어나므로 이또한 하지 않겠습니다.

#----------------

#해당 데이터셋에 케이스가 102개밖에 되지 않고 그마저도 위,경도가 포함된 데이터가 얼마 되지 않으므로 따로 서브셋시키지 않겠습니다.

```
m <- leaflet() %>%  #파이프함수를 사용하여 leaflet,addTiles,addMarkers 함수를 한번에 실행시켜줍니다.
  addTiles() %>%
  addMarkers(lng=covid_case$long, lat=covid_case$lat, popup=covid_case$name,)
```

```
## Warning in validateCoords(lng, lat, funcName): Data contains 71 rows with either
## missing or invalid lat/lon values and will be ignored
```

```
m
```