[패키지 정리]

pyreadstat 패키지 : pandas 패키지 함수를 이용해 SPSS, SAS, STATA 등 다양한 통계 분석

프트웨어의 데이터 파일을 불러올 수 있음

KoNLPy 패키지 : 한글 텍스트로 형태소 분석할 수 있는 패키지

re 패키지 : 문자 처리 패키지

wordcloud 패키지 : 워드클라우드 만드는 패키지

Cf) spss : IBM이 주인인 통계 프로그램 // 사회과학용 통계 패키지

<chapter 9>

**(데이터프레임명).dtypes** : 데이터프레임의 변수 타입을 담은 어트리뷰트

**막대그래프 정렬 순서 변경하기** : order 이용

Ex) sns.barplot(data=ageg\_income, x = ‘ageg’, y = ‘mean\_income’, order = [‘young’, ‘middle’, ‘old’])

# 사용할 데이터는 ageg\_income, x축은 ageg, y축은 mean\_income이고,

# 정렬 순서는 young, middle, old 순서로 막대그래프 정렬

**비율 구하기 : value\_counts( )에 normalize = True 입력**

Ex) rel\_div = welfare.query(‘marriage != “etc”’) \ #marriage에서 etc 제외

.groupby(‘religion’, as\_index = False) \ #religion별 분리

[‘marriage’] \ #marriage 추출

.value\_counts(normalize = True) #비율 구하기

**(데이터프레임명).round(출력할 자릿수)** : 값을 반올림하기

Ex) rel\_div = rel\_div.query(‘marriage == “divorce”’) \ #divorce추출

.assign(proportion = rel\_div[‘proportion’]\*100 \ #백분율로 바꾸기

.round(1) #반올림해서 소수점 첫째자리까지 출력

**피벗** : 행과 열을 회전해서 표의 구성을 바꾸는 작업

**(데이터프레임명).pivot(**

**index = ‘회전 기준 변수명’,**

**columns = ‘열로 구성할 변수명’,**

**values = ‘항목 값으로 구성할 변수명’)** : 데이터프레임을 피벗하기

ex) pivot\_df =

region\_ageg[[‘region’,’ageg’,’proportion’]]

.pivot(index = ‘region’, columns = ‘ageg’, values = ‘proportion’)

# region\_ageg데이터프레임의 region, ageg, proportion 변수 추출

# 지역(region)을 기준으로 데이터프레임 회전

# 연령대(ageg)별로 열을 구성하도록

# 각 항목 값을 비율(proportion)로 채우도록

**(데이터프레임명).plot.barh( )** : 가로 막대 그래프 만들기

+) **막대 그래프 누적해서 그릴 시, ( ) 안에 stacked = True 입력**

Ex) pivot\_df.plot.barh(stacked = True) #누적해서 가로 막대 그래프 만들기

<chapter10>

**re.sub( )** : str변수일 때, 불필요한 문자(특수문자/한자/공백 등) 제거하기

Ex) moon = re.sub(‘[^가-힣]’, ‘ ‘, moon) #moon에서 불필요한 문자 제거

**Cf) [^가-힣] : 한글이 아닌 모든 문자를 의미하는 정규표현식**

**str.replace(** **)** : 데이터 프레임에 담겨있는 변수일 때, 불필요한 문자(특수문자/한자/공백 등) 제거하기

Ex) df[‘reply’] = df[‘reply’].str.replace(‘[가-힣]’, ‘ ‘, regex = True)

#df의 reply 변수에서 불필요한 문자 제거

**konlpy.hannanum.nouns( )**

: 문장에서 명사를 찾아서 출력해 줌

: 출력 결과는 리스트 자료 구조임

Ex) hannanum.nouns(moon) #moon에서 명사만 찾아서 출력

**(데이터프레임명).str.len( )** : 단어의 글자수를 알려줌 (pandas내의 함수)

**워드클라우드**

: 단어의 빈도를 구름 모양으로 표현한 그래프

: 단어의 빈도에 따라 글자의 크기나 색깔이 다르게 표현되므로 어떤 단어가 얼마나 많이 사용됐는지 한눈에 파악 가능

**wc.generate\_from\_frequencies( )** : 워드클라우드 만들기

**plt.imshow( )** :워드클라우드 출력하기 (matplotlib.pyplot 패키지 내의 함수)

**apply( )** : 데이터 변환을 해 주는 함수 (pandas 내의 함수)

**(데이터프레임명).explode( )** :한 행에 한 단어만 들어가도록 하는 함수

Ex) nouns = nouns.explode( ) #nouns 자료를, 한 행에 한 단어만 들어가도록 설정