~~영화관람의 자유~~ _ 1912168 _ 이호림 제출파일

중요도 □□□ 재복습 1st 2nd 3rd

<R데분 함수 총 정리> 기말본에서까지

데이터파일이름 <- read.csv("파일명.csv", stringsAsFactors=F)

head(데이터, k) : 데이터를 위에서 k행까지

tail(데이터, k) : 데이터를 아래에서 k행까지

View(데이터) : 데이터를 보여줌 (Data 창에서 바로 클릭도 ok)

dim(데이터) : 행과 열이 각각 몇개씩 있는지 알려줌 (예문과)

str(데이터 or 데이터$변수) : 해당값의 속성을 보여줌 (범주인지 문자인지)

데이터$변수 <- NULL : 해당변수삭제

rm(데이터) : Data창에서 데이터 삭제

데이터$변수 <- as.factor(데이터$변수) : 해당변수의 형식을 범주형으로 변경 (문자: chr, 범주: factor)

summary(데이터) : 해당데이터의 기초통계량을 요약

mean(데이터$변수) : 해당변수의 평균

var(데이터$변수) : 해당변수의 분산 (분산이 클수록 편차가 큼)

sd(데이터$변수) : 해당변수의 표준편차

round(연산값, digits=n) : 해당값에서 소수 n번째 자리까지

└ 함수를 이용한 값가능 ex) mean(exam$math)

hist(데이터$변수, breaks=seq(N1, N2, by=k)) : 해당변수를 N1~N2까지 간격을 k인 히스토그램 그리기 ** 변수의 형식이 숫자형 / N1과 N2는 summary(데이터$변수)통해최소,최대

table(데이터$변수) : 해당변수의 빈도수 (형식이 문자나 범주인변수만 가능)

table(데이터$변수, 데이터변수2) : 두변수를 조합한 빈도수

---

install.packages("패키지") : 패키지를 설치함

library(패키지) : 설치된 패키지를 불러옴

⊕ remove.packages("패키지 이름") 을 통해서도

** 위전 ggplot2 패키지 불러오기 library(ggplot2)

qplot(data=데이터 변수) : 해당변수의 빈도를 막대그래프로 나타냄 my story 자이스토리

└ 문자형이여도 " " 안써도됨

색상위로 ←

qplot(data=데이터, 변수1, fill=변수2) : 변수1과 변수2의 조합된 빈도를 막대그래프로 표현

중요도 □ 1st 재복습 2nd 3rd

데이터$변수 <— as.Date(데이터$변수) : 해당변수의 형조를 날짜(Date)로 변경

weekdays(데이터$변수): 해당변수(날짜)의 요일이 무엇인지 알려줌

└ 이 변수의 형조가 날짜(Date)여야함.

weekdays(as.Date("2020-01-01")) : "2020-01-01"을 날짜(Date)로 인식한뒤 이 날짜의 요일을 알려줌

데이터$(A) <—— B. : B라는 값을(함수조합성) 데이터의 새로운 변수 'A'에 저장

└새변수 └아니나 여러데이터의 변수여도ok ex)데이터$변수

mean/var/sd(데이터$변수, na.rm=T) : 데이터 변수에서 결측치(NA)제외하고 계산

데이터$변수 <— factor(데이터$변수, levels=c("a","b","c"...))

: a, b, c 순으로 변수를 배열하고 형조를 범주형으로 바뀐뒤 해당데이터 변수에 지정

Stat.desc(데이터$변수): 해당변수의 여러 기술통계량 보여줌 % summary

★★ └우선 'pastecs' 설치하고 불러오기

describe(데이터$변수) % Stat.desc

└우선 'Psych' 설치하고 불러오기

★★★ 우선 "readxl" 패키지 설치하고 불러오기 ★★★

데이터파일이름 <— read_excel("파일명.xlsx", sheet=k) : 엑셀파일에서 k번째 시트에서 불러오기

write.csv(보낼때 저장할이름, file="파일명.csv") : B을 A라는 이름으로 내보내기 (csv형식으로)

A B

★★ 우선 "writexl" 패키지 설치하고 불러오기 ★★

write_xlsx(보낼때 저장할이름, path="파일명.xlsx") : B을 A라는 이름으로 내보내기 (excel형식으로)

A B

★★★

예를들어 welfare에있는 education변수을 welfare_7에 복사하시오

1) welfare_7$education <— welfare$education

2) welfare_7 <— welfare_7 %>% mutate(education = welfare$education)

my story 자이스토리

중요도 □ 1st 재복습 2nd 3rd

table(데이터$변수==k) : 해당변수가 k인갯수(문자형태면 " ")
table(데이터$변수 != k) : 해당변수가 k가아닌 갯수
table(데이터$변수 > k) : 해당변수가 k보다 큰갯수
< k : 해당변수가 k보다 작은 갯수
≥ k : 해당변수가 k보다 같거나큰갯수
≤ k : 해당변수가 k보다 같거나 작은갯수

table(데이터$변수 + 비교연산자 & 데이터$변수 + 비교연산자) / | (조가나 또는)
~이고, ~이면서 ~이거나, 또는
==/!/≥≤>< 다가능

: 해당 조건을 만족하는 갯수
table(데이터$변수 %in% c("a","b","c...));
데이터 변수가 a이거나 b이거나 c인 것들의 갯수 (|연산자 사용과 동일)
→ %in% c(~) : ~인애들을 다 동등

★★우선 "dplyr" 패키지 설치하고 불러들이★★
1) 데이터 <- rename(데이터, 바꿀변수명 = 이전변수명) — 여러변수동시에 바꾸기가능
2) 데이터 <- 데이터 %>% rename(바꿀변수명=이전변수명)
→변수가 문자형이어도 " " 안써도 ok.

데이터 <- 데이터(예전 gg plot2 설치하고 이 패키지에 들어있는 데이터만가능)
데이터 <- gg plot2 :: 데이터 : 데이터의 원본을 가져오기(단 gg plot2패키지 안에 있는 데이터만가능) ★★ "gg plot2" 패키지 설치하고 불러오기(ex) mpg, mtcars)
데이터$변수 <- ifelse(데이터$변수 == "a", "b", 데이터$변수) (ifelse두번씀)
: 만약 데이터의 변수값이 a이면 b로 바꾸고 아니면 그대로 데이터의 원래 변수명 그대로 두어라.
table(is.na(데이터$변수)) : 데이터의 해당변수의 NA(결측치)가 몇개인지
table(!is.na(데이터$변수)) : 데이터의 해당변수에서 NA(결측치)가 아닌게 몇개인지

ⓐ 예를들어서 데이터 프레임 weather_new의 변수 평균기압이 NA일경우는 ★★
is.na(weather_new$평균기압)이지
weather_new$평균기압==NA는 오답
%>% is.na(데이터$변수) : 해당변수가 NA이다

my story 자이스토리

filter 함수안에 다양한 여러 변수들을 기준으로 조건에 맞게 설정가능
ex) filter(class != 4, math >= 90 | history >= 95) : 4반이아니면서 수학이 90이상이거나 역사가 95이상

중요도 □ 재복습 2nd 3rd

★★★★★ - " " 안써도됨

데이터전처리는 (우선 'dplyr' 패키지 설치후 불러오기 ★★)

데이터 %>% filter(변수 == 사례) : 데이터 안에 있는 변수에서 해당사례만 추출

⊕ var나 mean과 같은 기술통계함수를 이용하기위해서는 해당사례를 만족하는것들을 새로운 데이터파일에 assign해야함

데이터 %>% filter(변수 >= quantile(변수, probs = c(k))) : 변수의 값에서 k인값

.데이터의 해당변수의값이 변수의 값들에서 k인 값보다 이상인사례만 추출

즉 k가 0.9이면 0.9 이상인값, 변수가 상위 10% 인사례들만추출(변수의값이 상위 1-k 인값들만추출)

데이터 %>% select(변수) : 해당변수들만추출 (여러변수를 , 로연결해서 추출가능)

데이터 %>% select(-변수) : 해당변수를 제외하고 추출

데이터 %>% select(contains("특정단어")) : 특정단어가 포함된변수추출

데이터 %>% arrange(변수) 변수에 대한 오름차순정렬 ⊕두 변수를 같이 쓸수있음 arrange(class, -math)

데이터 %>% arrange(-변수) : 변수에 대한 내림차순정렬

└⊕ head(k) : 정렬된 사례들중에서 위에서 k까지

→ 이변수는 웬만큼 변수인상태만 가능 ex) exam$math

이때라 math이나 -math인형태

데이터 ← 데이터 %>% mutate(새변수 = ☐) → 기존변수활용 or 다른데이터이용가능

☐ 라는 과정을 거친 새변수를 데이터에 지정, mutate함수로 동시에 여러변수 생성가능

(단순값들만 보고싶으면 데이터파일에 assign 안해도됨)

☐ 과정에 ifelse 함수가 쓰였는데 ifelse 안에 또 ifelse 가들어갈수있음

데이터 %>% group_by(변수) %>% summarise(mean or varorsd ...)

: 변수별로 그룹을 나누고 기술통계량을 조건에 맞게 제시 → 두 개의 변수 조합가능

: summarise를 붙여주는 결과값앞에 변수명지정가능 ex) mean_math = mean(math)

★★★★★ mutate든 summarise를이용하든 count = n() 빈도이용한 코드 ★★

count / sum(count) : 해당변수들끼리 나뉘어진 사례들끼리의 전체 갯수에서 해당사례비율

my story 자이스토리

count / sum(데이터$count) : 모든사례들(전체) 갯수에서 해당사례 비율

print(n=k) : k줄까지 다 보여줌(맨끝까지 다 보고싶으면 복잡하니까 print=100)

중요도 □ □ □ 재복습 1st 2nd 3rd

데이터 <— 데이터 %>% relocate (변수, . before = 변수2) : 변수를 변수2 왼쪽에

데이터 <— 데이터 %>% relocate (변수, . after = 변수2) : 변수를 변수2 오른쪽에

데이터 <— left_join (데이터, 데이터2, by = "변수") : 데이터와 데이터2를 공통변수를 이용해 합침 (데이터에 저장)

데이터프레임명 <— data.frame (변수1 = c (변수1의사례나열), 변수2 = c (변수2의사례나열)) : 변수1과 변수2를 갖는 새로운 데이터 프레임생성 (변수가 문자형이면 " ")

★★ - 사례들이 단순 숫자면 : 이용해서 표시 ex) 1:7 (단 1~7연목일때)

→ (공통변수의 척도가 동일한게 좋음 str( ) 이용해서 as.factor( )로 척도를 범주형으로 일치)

데이터 <— left_join (데이터, 데이터2, by = c ("변수1" = "변수2"))

: 공통변수가 없을때 변수1과 변수2가 같다고 지정한뒤 데이터들끼리 합치기

데이터 <— bind_rows (데이터, 데이터2) : 데이터와 데이터2를 합쳐 사례추가

데이터 <— 데이터 %>% distinct (변수, . keep_all = T) :

해당변수가 같은 사례들중에 한개만 남기고 다 제외하기 → 변수최대3개가 적당 (값이 같은변수만)

→ (exam에 3개만같아서 자위진 사례도 있으니 안전빵으로 4개 적기)

★★★ mIdwest 실습은 우선 'ggplot2' 패키지 설치하고 불러오기 ★★

데이터 <— ggplot2 :: 데이터 이용해서 파일 불러오기 후 실습

데이터 <— 데이터 %>% drop_na( ) : 데이터에서 다양한변수들중에서 그값이 단 하나라도 NA가있다면 그 사례는 싹 다제거 → 우선 'tidyr' 패키지 설치후불러오기

데이터 %>% select (A : last_col ( )) : 데이터에서 A부터 마지막변수까지추출

★★ sav 형태의 파일은 'foreign' 패키지 설치후 불러오기

데이터 <— read.spss (file = "처음저장할당시파일이름", to.data.frame = T)

: spss 파일불러서 데이터 프레임 만들기 └ 파일명.sav 형태

ggplot (데이터, aes (변수1, 변수2, fill = 변수2)) + geom_bar (stat = "identity")

: 변수1과 변수2를 각각 X축 Y축이되고 '값'들을 나타내는 막대그래프만들기

'ggplot2' 패키지불러오기필수

중요도 □ □ □ 재복습 1st 2nd 3rd

★★★ 우선 library(ggplot2) 하기 ★★★

ggplot(데이터, aes(reorder(변수1, 변수2), 변수2, fill= 변수2))
+ geom_bar(stat="identity") + theme(axis.text.x = element_text(size=7.5, angle=50))
: 변수1과 변수2를 표현하는 막대 그래프 만들기

새로운데이터 <— 기존데이터 : 기존데이터의 복사본인새로운데이터 만들기
A:<— C:(1:30) : 1부터 30까지의 A라는 이름의 열벡터 생성

colors() : 색깔을 나타내줌 (그래프나 텍스트마이닝 이용)