```
##############################################
### TextMining Lecture 02                 #####
### Subject: RPython Programming          #####
### Developed by. Dr. Keungoui KIM        #####
### https://awekim.github.io/portfolio/ #####
##############################################

### Help
?print

### Quick overview
data(mtcars)

a <- 1
a

myvec <- 1:10
mydf <- mtcars[1:10,]
mylist <- list(myvec, mydf)

myvec[c(1,3)]
mydf$mpg
length(mydf)
length(mylist)
mylist[[2]]$mpg

### Load file
load(file="R file/R file_LEC02/ds_salaries_ed.RData")

getwd()

### dplyr
sdsdsd <- 1
ds_sal %>% head
library(dplyr)
ds_sal %<>% head
head(ds_sal)

library(magrittr)

### select
# column name
ds_sal[,c("job_title","salary","salary_currency")]
```

```
head(ds_sal[,c("job_title","salary","salary_currency")])

library(dplyr)
ds_sal %>% #names
  select(job_title, salary, salary_currency) %>%
  head

ds_sal %>% select(job_title:salary_currency) %>%
  head

# index
head(ds_sal[,c(5,7)])

ds_sal %>% select(5,7) %>%
  head

ds_sal %>% select(5:7) %>%
  head

# select with starts_with
ds_sal %>% #names
  select(salary) %>%
  head

ds_sal %>% select(starts_with('salary')) %>%
  head

ds_sal %>% select(!starts_with('salary')) %>%
  head

# ! Question
ds_sal %>%
  select(starts_with('salary') | starts_with('company')) %>%
  names

# select with if
ds_sal %>% head(1)

ds_sal %>%
  select_if(is.numeric) %>%
  head(2)

ds_sal %>%
```

```
  select_if(is.character) %>%
  head(2)


### Filter
head(ds_sal[ds_sal$job_title=="Data Scientist",], 2)

ds_sal %>% head(1)
ds_sal %>%
  select(job_title) %>%
  unique
ds_sal$job_title %>% unique

ds_sal %>%
  filter(job_title=="Data Scientist") %>%
  head(2)

ds_sal %>%
  filter(job_title=="Data Scientist") %>%
  select(job_title) %>% unique

head(ds_sal[ds_sal$salary>=mean(ds_sal$salary),], 2)

mean(ds_sal$salary)
ds_sal %>%
  filter(salary>=mean(salary)) %>%
  head(2)

### REVIEW
# What is the highest salary among those working
# for large corporations?
ds_sal %>%
  filter(company_size=='L') %>%
  arrange(desc(salary)) %>%
  head(3)

# What is the average salary of people who are working
# fully remotely?
# Answer this questions with two versions: conventional
approach & chain operator approach
mean(ds_sal[ds_sal$remote_ratio==100,]$salary)

ds_sal %>%
```

```
  filter(remote_ratio==100) %>%
  select(salary) %>%
  pull() %>%
  mean()

# Where do the top 10 highest-earning individuals live?
ds_sal %>%
  arrange(desc(salary)) %>%
  select(employee_residence) %>%
  head(10)

# Is it possible to be a Data Scientist who works
# full time (FT),
# fully remotely for the large company?
# If possible, how many cases are there?
ds_sal %>%
  filter(job_title=="Data Scientist" &
           employment_type=='FT' &
           remote_ratio==100 &
           company_size=='L') # %>% nrow

### arrange
head(ds_sal[order(ds_sal$salary),],2)

ds_sal %>%
  arrange(salary) %>%
  head(2)

head(ds_sal[order(ds_sal$salary, decreasing=TRUE),],2)

ds_sal %>%
  arrange(desc(salary)) %>%
  head(2)

### mutate
head(ds_sal, 2)

ds_sal %>% mutate(experience=2024-work_year) %>%
  select(work_year,experience,salary) %>%
  head

ds_sal %>%
  mutate(salary.d = ifelse(salary_in_usd > mean(salary_in_usd),
```

```
                              "High", "Low")) %>%
  select(work_year,salary,salary.d) %>% head

# ! Question
ds_sal %>% mutate(International =
                    ifelse(employee_residence !=
company_location,
                              "International","Domestic")) %>%
  select(employee_residence,company_location,International) %>%
  head

ds_sal %>%
  mutate(job.d = case_when(job_title=="Data Scientist" ~ "DS",
                              job_title=="Data Analyst" ~ "DA",
                              TRUE ~ "Others")) %>%
  select(work_year,job_title,job.d) %>% head

ds_sal %<>%
  mutate(experience=2023-work_year) %>%
  mutate(salary.d = ifelse(salary_in_usd > mean(salary_in_usd),
                              "High", "Low")) %>%
  mutate(job.d = case_when(job_title=="Data Scientist" ~ "DS",
                              job_title=="Data Analyst" ~ "DA",
                              TRUE ~ "Others"))

# mutate_at
ds_sal %>% select(ID, salary,experience) %>%
  mutate_at(vars(salary,experience), log) %>% head

ds_sal %>% select(ID, salary,experience) %>%
  mutate_at(vars(salary,experience), max) %>% head



# mutate_all
ds_sal %>%
  mutate_all(is.na) %>% head(2)

norm.fun <-
  function(x){
    (x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)}
ds_sal %>% select_if(is.numeric) %>%
  mutate_all(norm.fun) %>% head
```

```
ds_sal.1 <- ds_sal %>%
  rbind(work_year=NA, salary=NA,
        salary_in_usd=NA,
        remote_ratio=NA)
ds_sal.1 %>% select_if(is.numeric) %>%
  mutate_all(norm.fun) %>% head

### rename
names(ds_sal)
ds_sal %<>% rename(sal.type=salary.d,
                   job.type=job.d)
names(ds_sal)

ds_sal %>% rename_with(toupper) %>% names
ds_sal %>% rename_with(toupper, starts_with("salary")) %>%
names

### group_by
ds_sal %>% class
ds_sal_gr <- ds_sal %>%
  group_by(job_title)
ds_sal_gr %>% class
ds_sal_gr %>% ungroup

ds_sal_gr %>% class
ds_sal_gr %>% ungroup %>% class
ds_sal_gr %>% data.frame %>% class

### group_by + mutate
ds_sal %>%
  # group_by(company_size) %>%
  mutate(salary.mean=mean(salary)) %>%
  data.frame %>% head(2)

ds_sal %>%
  group_by(company_size) %>%
  mutate(salary.mean=mean(salary)) %>%
  data.frame %>% head(2)

### group_by + summarise
ds_sal_gr %>% summarise(salary=mean(salary))
ds_sal_gr %>% ungroup %>%
  summarise(salary=mean(salary))
```

```
ds_sal %>%
  group_by(job_title,company_size) %>%
  dplyr::summarise(salary=mean(salary))

### ggplot
library(ggplot2)

ds_sal %>% head

# scatter plot
plot(ds_sal$experience, ds_sal$salary)
ds_sal %>% ggplot(aes(x=experience, y=salary)) +
  geom_point()

ds_sal %>%
  ggplot(aes(x=experience, y=salary)) +
  geom_point(color="red")
ds_sal %>%
  ggplot(aes(x=experience, y=salary,
             color=experience_level)) +
  geom_point()

# line plot
ds_sal %>% group_by(work_year) %>%
  summarise(salary_in_usd = mean(salary_in_usd)) %>%
  ggplot(aes(x=work_year, y=salary_in_usd)) +
  geom_line()

library('ggplot2')
ds_sal %>%
  group_by(work_year, experience_level) %>%
  summarise(salary_in_usd = mean(salary_in_usd)) %>%
  ggplot(aes(x=work_year, y=salary_in_usd,
             group=experience_level,
             color=experience_level)) +
  geom_line()

# bar plot
ds_sal %>% ggplot(aes(company_size)) +
  geom_bar()

ds_sal %>% ggplot(aes(job_title)) +
```

```
  geom_bar() + coord_flip()

ds_sal %>%
  ggplot(aes(x=job_title, fill=company_size)) +
  geom_bar(position='fill') + coord_flip()

ds_sal %>% names
ds_sal %>% group_by(job.type, company_size) %>%
  summarise(salary=mean(salary)) %>%
  ggplot(aes(x=job.type, y=salary, fill=company_size)) +
  geom_bar(stat='identity')

ds_sal %>% group_by(job.type, company_size) %>%
  summarise(salary=mean(salary)) %>%
  ggplot(aes(x=job.type, y=salary, fill=company_size)) +
  geom_bar(stat='identity', position='dodge')
```