Lab 2 - In-Class Exercise

Data Visualization DS4073

In-class exercise

Please download the `Salaries_dirty.csv` from iSpace and complete the following requirements.

```
# loading package
library(dplyr)
library(readr)
```

```
# 1. Please read Salaries data from the file Salaries_dirty.csv
# write you code here
data <- read_csv("Salaries_dirty.csv")
head(data)
```

```
## # A tibble: 6 x 6
##   rank      discipline yrs.since.phd yrs.service sex   salary
##   <chr>     <chr>              <dbl>       <dbl> <chr>  <dbl>
## 1 Prof      B                     19          18 Male  139750
## 2 Prof      B                     20          16 Male  173200
## 3 AsstProf  B                      4           3 Male   79750
## 4 Prof      B                     45          39 Male  115000
## 5 Prof      B                     40          41 Male  141500
## 6 AssocProf B                      6           6 Male   97000
```

```
# 2. Select the female professors (including assistant professor, associate professor, and full profess

# write you code here
prof <- filter(data, sex == "Female" & salary > 50000 & salary < 100000 & rank == "AsstProf" || rank ==
head(prof)
```

```
## # A tibble: 6 x 6
##   rank      discipline yrs.since.phd yrs.service sex   salary
##   <chr>     <chr>              <dbl>       <dbl> <chr>  <dbl>
## 1 Prof      B                     19          18 Male  139750
## 2 Prof      B                     20          16 Male  173200
## 3 AsstProf  B                      4           3 Male   79750
## 4 Prof      B                     45          39 Male  115000
## 5 Prof      B                     40          41 Male  141500
## 6 AssocProf B                      6           6 Male   97000
```

```
# 3. Calculate the mean of income of professors (all types of professor) grouped by sex
```

```r
# write you code here
res <- data %>%
  group_by(sex) %>%
  summarize(mean_income = mean(salary, na.rm=TRUE))
print(res)
```

```
## # A tibble: 2 x 2
##   sex    mean_income
##   <chr>        <dbl>
## 1 Female     101002.
## 2 Male       115186.
```

```r
# 4. There are some missing values in the dataset. Please calculate the proportion of the missing value

# write you code here
prop <- colSums(is.na(data)/nrow(data))
prop
```

```
##          rank    discipline yrs.since.phd   yrs.service           sex
##    0.00000000    0.00000000    0.01007557    0.01511335    0.00000000
##        salary
##    0.01511335
```

```r
# 5. Please impute the missing values with the 10 nearest neighbors and then calculate the mean of inco

library(VIM)

# write you code here
# Impute missing values using k-nearest neighbors
data_rmnull <- kNN(data, k = 10)

#2. calculate the mean of income of professors (all types of professor) grouped by sex
res <- data_rmnull %>%
  group_by(sex) %>%
  summarize(mean_income = mean(salary, na.rm=TRUE))
print(res)
```

```
## # A tibble: 2 x 2
##   sex    mean_income
##   <chr>        <dbl>
## 1 Female     101002.
## 2 Male       114992.
```