HW2.R

joann

2021-02-12

```r
#Missing Data Homework 2

#Read and inspect data set
data <- read.csv("C:/Users/joann/OneDrive/Desktop/missing data/week 1/missingdata_hw1.csv", na.strings =
str(data)
```

```
## 'data.frame':    2129 obs. of  14 variables:
##  $ enrollee_id           : int  32403 9858 31806 27385 27724 217 21465 27302 12994 16287 ...
##  $ city                  : chr  "city_41" "city_103" "city_21" "city_13" ...
##  $ city_development_index: num  0.827 0.92 0.624 0.827 0.92 0.899 0.624 0.92 0.878 0.624 ...
##  $ gender                : chr  "Male" "Female" "Male" "Male" ...
##  $ relevent_experience   : chr  "Has relevent experience" "Has relevent experience" "No relevent expe
##  $ enrolled_university   : chr  "Full time course" "no_enrollment" "no_enrollment" "no_enrollment" .
##  $ education_level       : chr  "Graduate" "Graduate" "High School" "Masters" ...
##  $ major_discipline      : chr  "STEM" "STEM" NA "STEM" ...
##  $ experience            : chr  "9" "5" "<1" "11" ...
##  $ company_size          : chr  "<10" NA NA "10/49" ...
##  $ company_type          : chr  NA "Pvt Ltd" "Pvt Ltd" "Pvt Ltd" ...
##  $ last_new_job          : chr  "1" "1" "never" "1" ...
##  $ training_hours        : int  21 98 15 39 72 12 11 81 2 4 ...
##  $ gender2               : int  0 1 0 0 0 0 NA 1 0 0 ...
```

```r
#Find variables with missing values
sapply(data, function(x) sum(is.na(x)))
```

```
##            enrollee_id                   city city_development_index
##                      0                      0                      0
##                 gender    relevent_experience    enrolled_university
##                    508                      0                     31
##        education_level       major_discipline             experience
##                     52                    312                      5
##           company_size           company_type           last_new_job
##                    622                    634                     40
##         training_hours                gender2
##                      0                    532
```

```r
#All variables with missing values are categorical
#I need to generate missing values for a continuous variable

#Generate missing values for training_hours depending on one variable
library(dplyr)
```

```
##
## 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
```

```r
data_new = select(data,'city_development_index','training_hours')
library(mice)
```

```
## Warning: package 'mice' was built under R version 4.0.3

##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```r
cont_cat = ampute(data_new,prop = 0.2,patterns=c(1,0),mech = "MAR")$amp
data['training_hours'] = cont_cat['training_hours']

#To keep it simple, I only use one categorical variable with missing variable for analysis
#The variable gender2 have missing value > 20%
data2 = select(data,'enrollee_id','city','city_development_index','training_hours','gender2','relevent_
#Check for missingness
sapply(data2, function(x) sum(is.na(x)))
```

```
##            enrollee_id                   city city_development_index
##                      0                      0                      0
##         training_hours                gender2    relevent_experience
##                    427                    532                      0
```

```r
#Mean imputation for numeric missing value
mean.imp <- function (a)
{
  missing <- is.na(a)
  a.obs <- a[!missing]
  imputed <- a
  imputed[missing] <- mean(a.obs)
  # Output the imputed vector
  return (imputed)
}
```

```r
data2['training_hours_imp']=mean.imp(data2['training_hours'])



#Mode imputation for the categorical variable
mode <- function (a)
{
  ta =table(a)
  tam = max(ta)
  if(all(ta==tam))
    mod =NA
  else
    mod = names(ta)[ta==tam]
  return (mod)
}

mode.imp <- function (a)
{
  missing <- is.na(a)
  a.obs <- a[!missing]
  imputed <- a
  imputed[missing] <- mode(a.obs)
  # Output the imputed vector
  return (imputed)
}

data2['gender_imp']=mode.imp(data2['gender2'])

#Analysis with complete case
data_complete = na.omit(data2)

#analyse the relationship between training hours and other factors

anova_one_way1 <- aov(training_hours~city_development_index+gender2,data=data_complete)
summary(anova_one_way1)
```

```
##                          Df  Sum Sq Mean Sq F value Pr(>F)
## city_development_index    1    1585    1585   0.446  0.504
## gender2                   1     927     927   0.261  0.609
## Residuals              1265 4491601    3551
```

```r
anova_one_way2 <- aov(training_hours_imp~city_development_index+gender_imp,data=data2)
summary(anova_one_way2)
```

```
##                          Df  Sum Sq Mean Sq F value Pr(>F)
## city_development_index    1    3059  3059.4   1.068  0.301
## gender_imp                1     581   580.8   0.203  0.653
## Residuals              2126 6088847  2864.0
```

```r
#For the imputed data, the sum of square for city development index is much larger and the sum of squar
#However, both variables does not pass the f test for both dataset
#The conclusion for both datasets is the same: we cannot reject the null hypothesis and different gende
```