MATH2349 Semester 1, 2018

Assignment 3

Student name(s) and numbers comes here

Required packages

Provide the packages required to reproduce the report. Make sure you fulfilled the minimum requirement #10.

```
library(readr)
library(tidyr)
library(dplyr)
library(Hmisc)
library(outliers)
library(magrittr)
library(plyr)
library(outliers)
library(knitr)
library(tidyverse)
library(GGally)
library(cowplot)
library(mlr)
setwd("C:/Users/lipy1/Desktop/Data processing/Assignment 3")
```

Executive Summary

The data is collected to predict incomes in 2 classifications, below 50,000 USD (<= 50k) or above 50,000 USD (> 50k) comparing incomes with variety of variables. The data is collected at the 1994 US, publicly sourced from the UCI Machine learning repository (http://archive.ics.uci.edu/ml/datasets/Census+Income (http://archive.ics.uci.edu/ml/datasets/Census+Income)). We first combine train sets and test sets then produce the data frame into meaningful form. And scanning properties of the data frame whether there exist NA or special values. Finally we inspect outliers of numerical columns by using multiple techniques to compare.

Data

```
train <- read.csv('adult.data.txt', header = FALSE)
test  <- read.csv('adult.test.txt', header = FALSE)

# Combine two sets of data.
adult <- rbind(train, test)

# Set headers
names(adult) <- c('age', 'workclass', 'fnlwght', 'education', 'education_num', 'marital_status', 'occupati
on', 'relationship', 'race', 'sex', 'capital_gain', 'capital_loss', 'hours_per_week', 'native_country', 'i
ncome')

# Delete white space at the beginning of characters
adult[, sapply( adult, is.character )] <- sapply( adult[, sapply( adult, is.character )], trimws)

# Delete row not in the format.
adult[c(32562),]
```

	age	workclass	fnlwght	education	education_num	marital_status
	<chr>	<fctr>	<int>	<fctr>	<int>	<fctr>
32562	|1x3 Cross validator		*NA*		*NA*	

1 row | 1-8 of 16 columns

```
adult <- adult[-c(32562),]

# Delete non-meaningful column
adult$fnlwght <- NA
```

Understand

```
adult$age <- as.numeric(adult$age)
adult$income <- factor(adult$income, ordered = TRUE)

str(adult)
```

```
## 'data.frame':    48842 obs. of  15 variables:
##  $ age           : num  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : Factor w/ 10 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ fnlwght       : logi  NA NA NA NA NA NA ...
##  $ education     : Factor w/ 17 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital_status: Factor w/ 8 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation    : Factor w/ 16 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ relationship  : Factor w/ 7 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
##  $ race          : Factor w/ 6 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex           : Factor w/ 3 levels " Female"," Male",..: 2 2 2 2 1 1 1 2 1 2 ...
##  $ capital_gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital_loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native_country: Factor w/ 43 levels " ?"," Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
##  $ income        : Ord.factor w/ 4 levels " <=50K"<" >50K"<..: 1 1 1 1 1 1 1 2 2 2 ...
```

Tidy & Manipulate Data I

Check if the data conforms the tidy data principles. If your data is not in a tidy format, reshape your data into a tidy format (minimum requirement #5). In addition to the R codes and outputs, explain everything that you do in this step.

```
adult
```

	…	workclass	fnlwght	education	education_num	marital_status
	<dbl>	<fctr>	<lgl>	<fctr>	<int>	<fctr>
1	39	State-gov	*NA*	Bachelors	13	Never-married
2	50	Self-emp-not-inc	*NA*	Bachelors	13	Married-civ-spouse
3	38	Private	*NA*	HS-grad	9	Divorced
4	53	Private	*NA*	11th	7	Married-civ-spouse
5	28	Private	*NA*	Bachelors	13	Married-civ-spouse
6	37	Private	*NA*	Masters	14	Married-civ-spouse
7	49	Private	*NA*	9th	5	Married-spouse-absent
8	52	Self-emp-not-inc	*NA*	HS-grad	9	Married-civ-spouse
9	31	Private	*NA*	Masters	14	Never-married
10	42	Private	*NA*	Bachelors	13	Married-civ-spouse

1-10 of 10,000 rows | 1-7 of 16 columns Previous **1** 2 3 4 5 6 … 1000 Next

```
# Data is already in tidy format. Variables/Attributes are columns and samples are rows.
```

Tidy & Manipulate Data II

```
# Average weekly capital
adult <- adult %>% mutate(Weekly = (capital_gain-capital_loss)/52)
```

Scan I

```
# Check NA data
colSums(is.na(adult))
```

```
##            age      workclass         fnlwght      education  education_num
##              0              0           48842              0              0
## marital_status     occupation   relationship           race            sex
##              0              0              0              0              0
##   capital_gain   capital_loss hours_per_week native_country         income
##              0              0              0              0              0
##         Weekly
##              0
```

```
# There is no NA values in the data frame.

# Special Values
is.special <- function(x){
if (is.numeric(x)) !is.finite(x) else is.na(x)
}

colSums(sapply(adult, is.special))
```

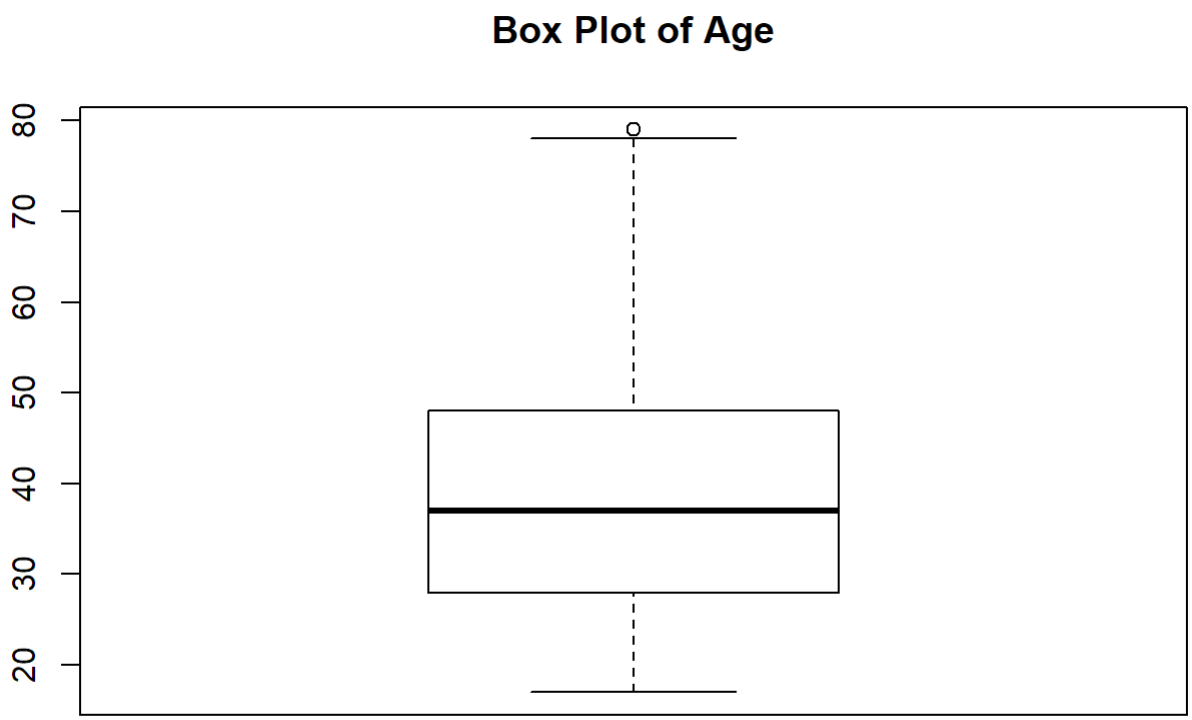
```
##            age      workclass         fnlwght      education  education_num
##              0              0           48842              0              0
## marital_status     occupation   relationship           race            sex
##              0              0              0              0              0
##   capital_gain   capital_loss hours_per_week native_country         income
##              0              0              0              0              0
##         Weekly
##              0
```

```
# There is no special values.
```

Scan II

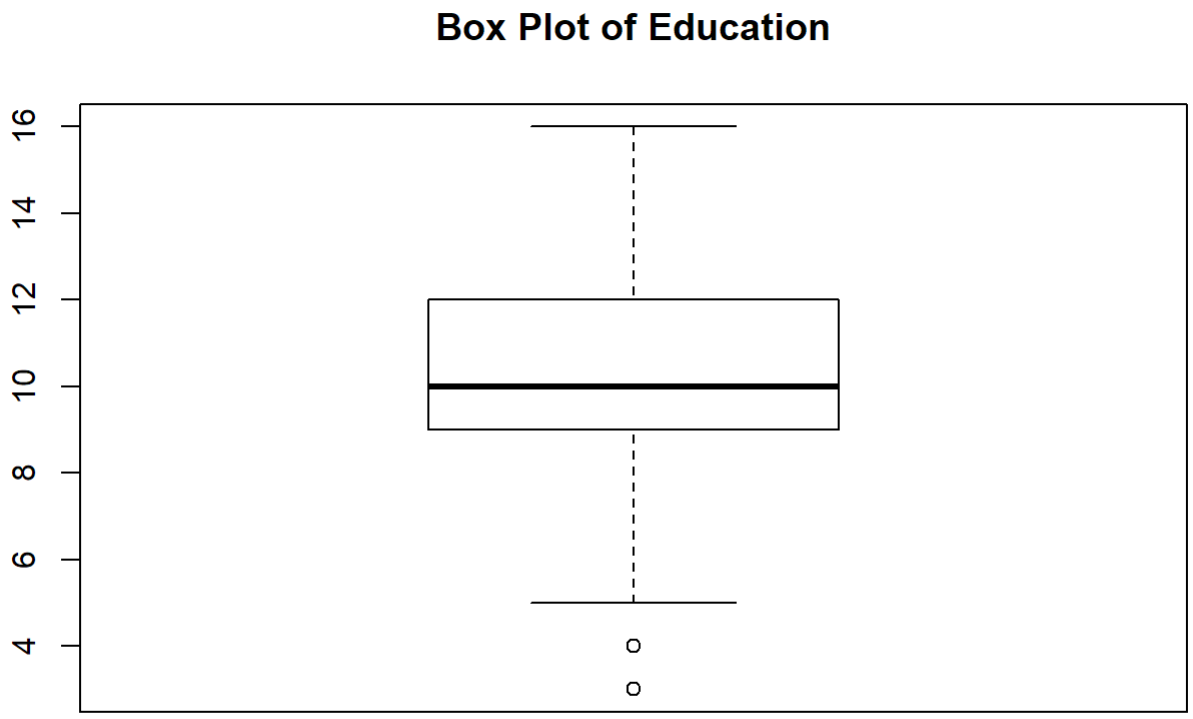
```
adult_clean <- adult

# Outlieer: Age
# Z-scores
z.scores <- adult_clean$age %>% scores(type = "z")
# Outliers
adult_out <- adult_clean$age[-which(abs(z.scores)>3)]
# Box Plot
adult_out %>% boxplot(main="Box Plot of Age")
```



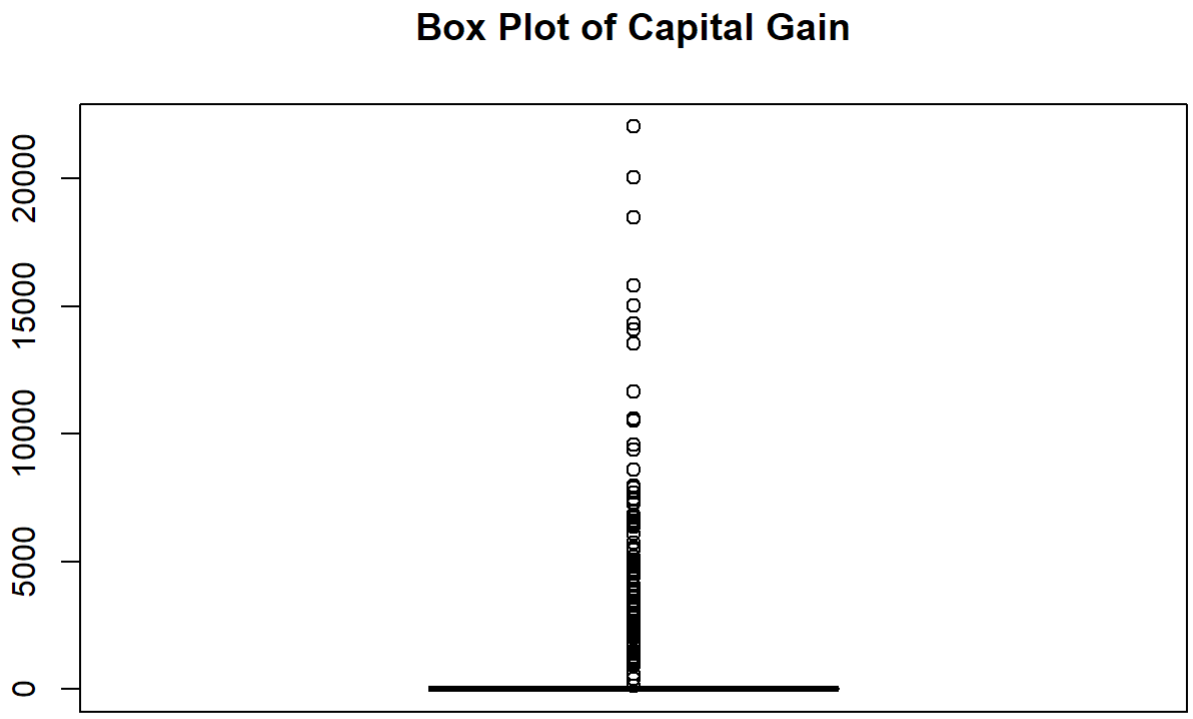

```
# Exclusion
adult_clean <- adult[-c(which(abs(z.scores)>3)),]
```

```
# Outlieer: education_num
# Z-scores
z.scores <- adult_clean$education_num %>% scores(type = "z")
# Outliers
adult_out <- adult_clean$education_num[-which(abs(z.scores)>3)]
# Box Plot
adult_out %>% boxplot(main="Box Plot of Education")
```



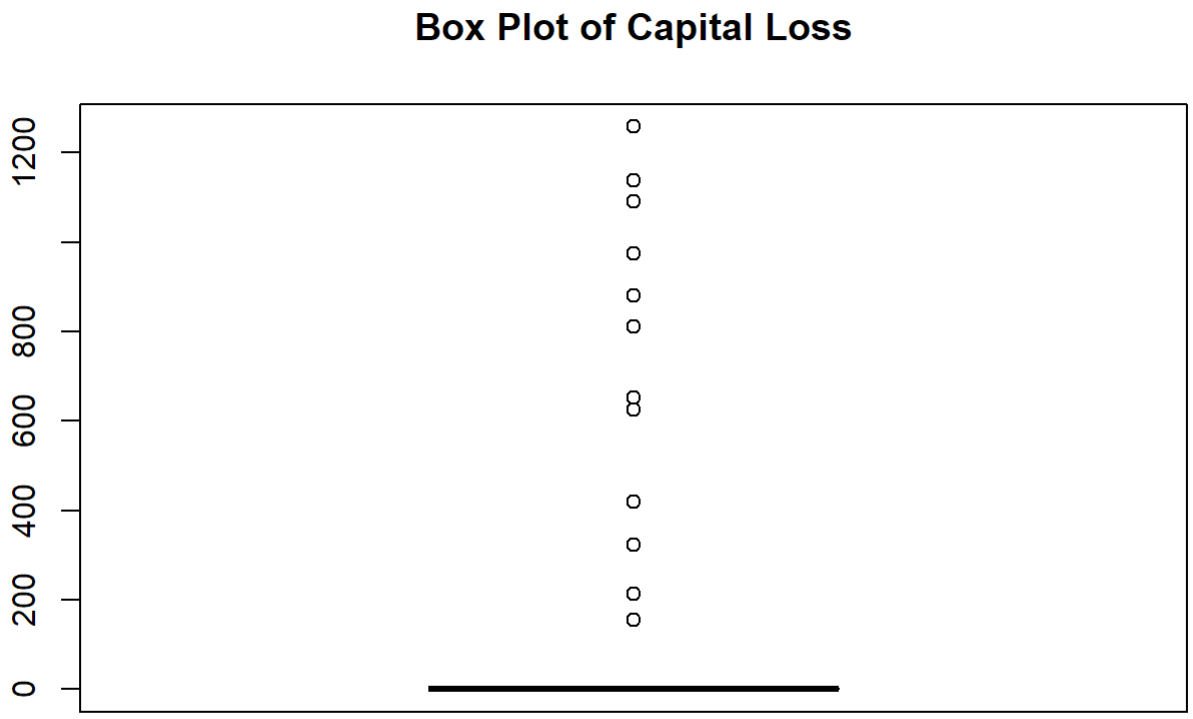

```
# Exclusion
adult_clean <- adult_clean[-c(which(abs(z.scores)>3)),]
```

```
# Outlieer: Capital Gain
# Z-scores
z.scores <- adult_clean$capital_gain %>% scores(type = "z")
# Outliers
adult_out <- adult_clean$capital_gain[-which(abs(z.scores)>3)]
# Box Plot
adult_out %>% boxplot(main="Box Plot of Capital Gain")
```



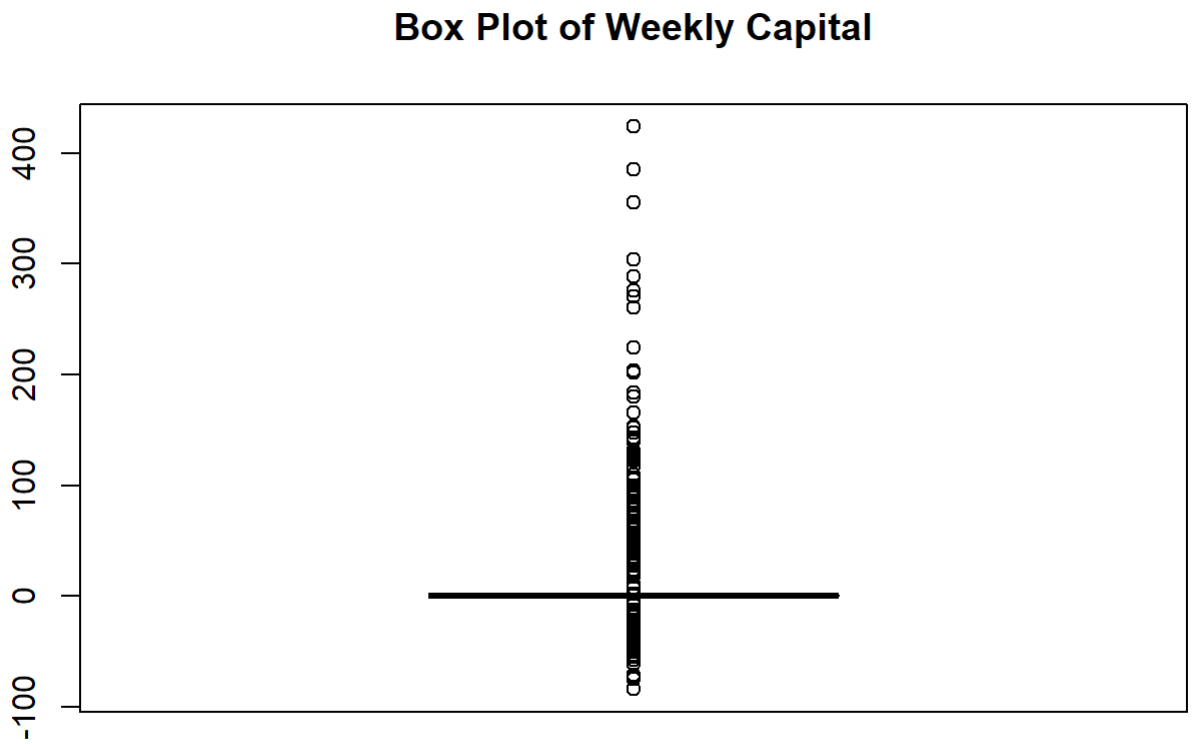

```
# Exclusion
adult_clean <- adult_clean[-c(which(abs(z.scores)>3)),]
```

```
# Outlieer: Capital Loss
# Z-scores
z.scores <- adult_clean$capital_loss %>% scores(type = "z")
# Outliers
adult_out <- adult_clean$capital_loss[-which(abs(z.scores)>3)]
# Box Plot
adult_out %>% boxplot(main="Box Plot of Capital Loss")
```




```
# Exclusion
adult_clean <- adult_clean[-c(which(abs(z.scores)>3)),]
```

```
# Outlieer: Weekly capital
# Z-scores
z.scores <- adult$Weekly %>% scores(type = "z")
# Outliers
adult_out <- adult$Weekly[-which(abs(z.scores)>3)]
# Box Plot
adult_out %>% boxplot(main="Box Plot of Weekly Capital")
```



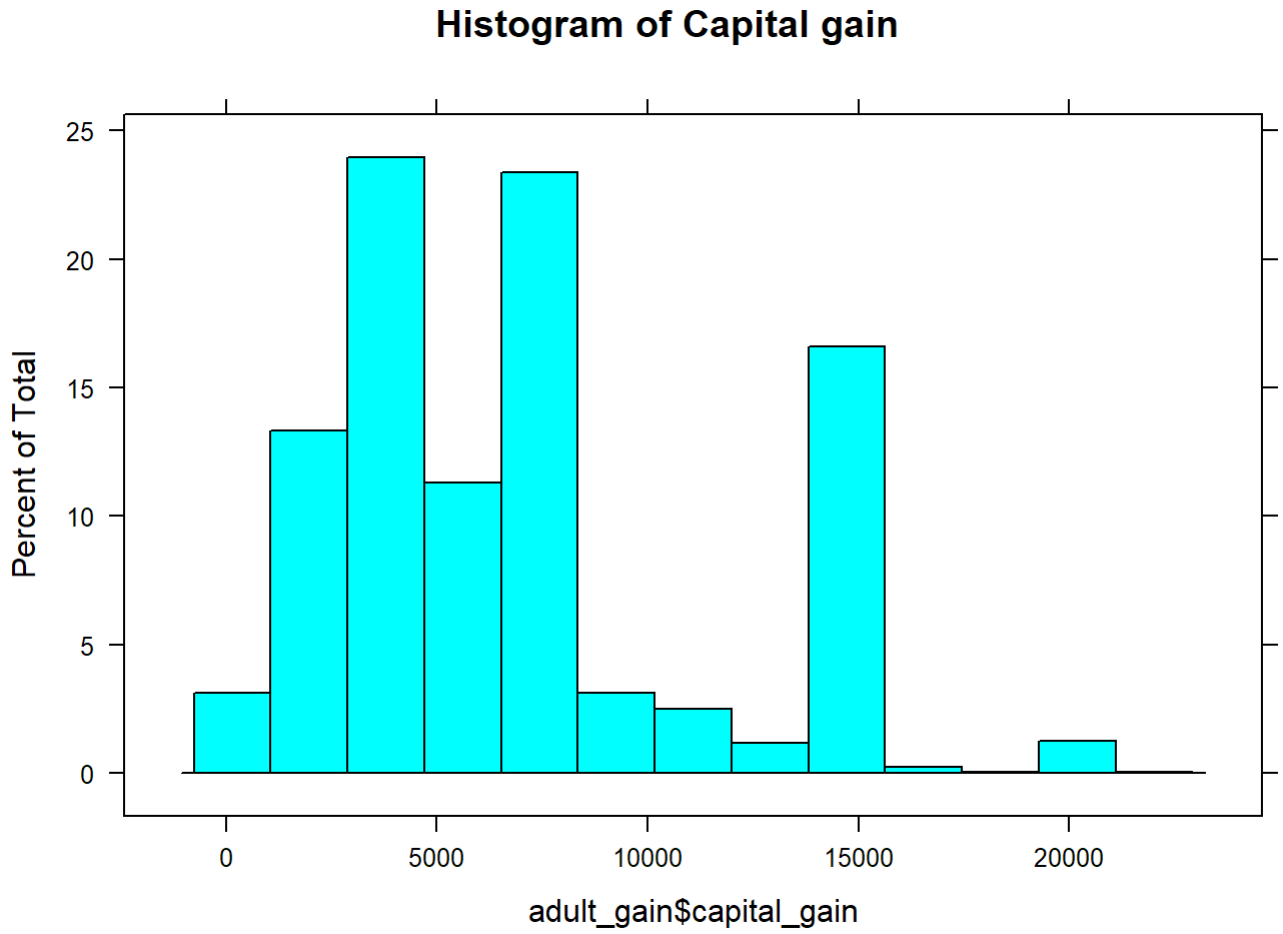

```
# Exclusion
adult_clean <- adult_clean[-c(which(abs(z.scores)>3)),]
```

Transform

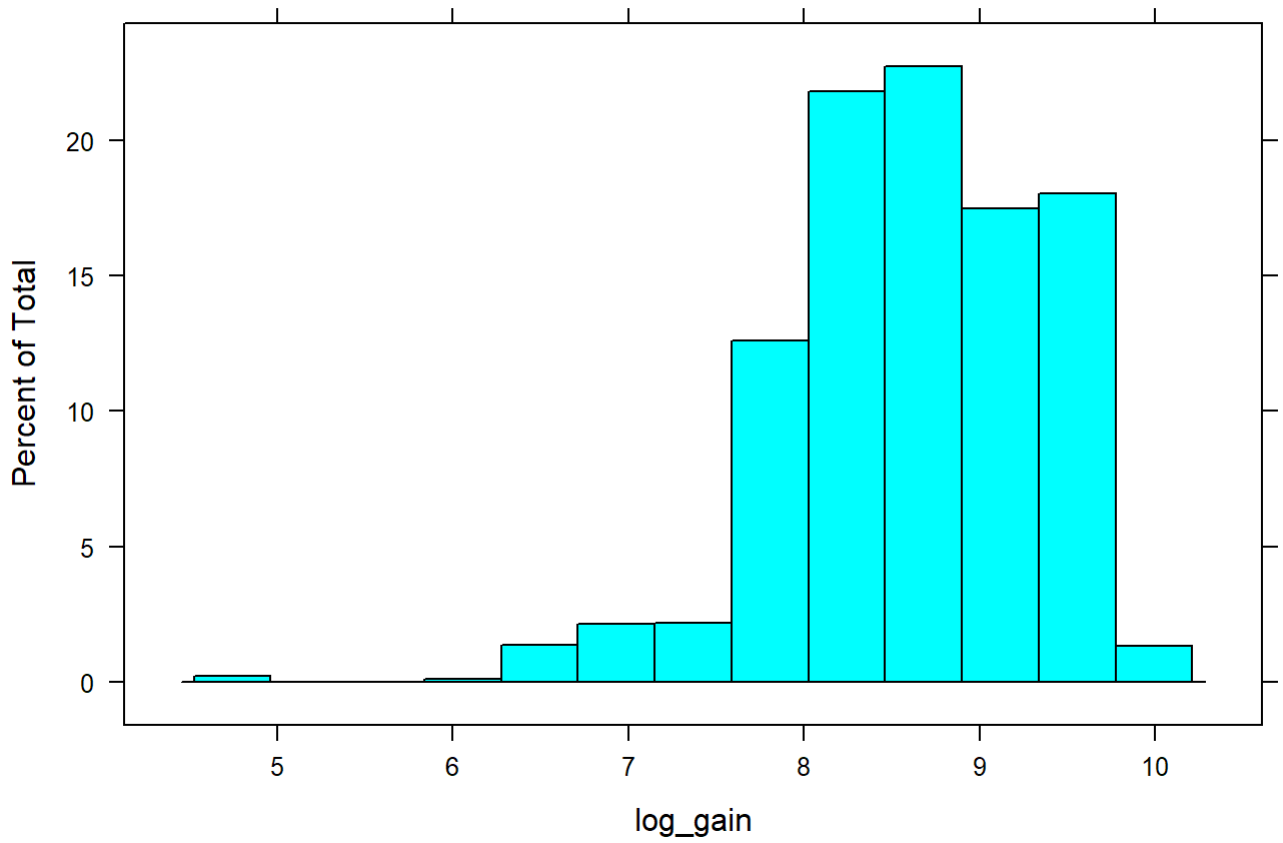
Apply an appropriate transformation for at least one of the variables. In addition to the R codes and outputs, explain everything that you do in this step. In this step, you should fulfil the minimum requirement #9.

```
# Exclude zero values from Capital gain attribute to get better scale of y.
adult_gain <- adult_clean[-which(adult_clean$capital_gain==0),]

# Histogram of Capital Gain
histogram(adult_gain$capital_gain, main="Histogram of Capital gain")
```




```
# It is hard to see the relation.
# Use Log-transformation to see linear relation.
log_gain <- log(adult_gain$capital_gain)
histogram(log_gain, main="Histogram of Log Capital gain")
```




```
# Now the capital gain can be told that it is Left-skewed.
```

```
# Further Outlier tests

# Chi-squared test
chisq.out.test(adult$capital_gain, variance=var(adult$capital_gain),opposite=FALSE)
```

```
##
##  chi-squared test for outlier
##
## data:  adult$capital_gain
## X-squared = 176.21, p-value < 2.2e-16
## alternative hypothesis: highest value 99999 is an outlier
```

```
# Extreme outlier in the capital gain attribute is $99999.

# By using outlier given in function, 99999 is a ouitlier as well.
outlier(adult$capital_gain)
```

```
## [1] 99999
```

```
# However, value 99999 is filtered from above z-test.
```