HW_1

Andus Kong; UNI: ak4479

9/17/2020

1. Form Project Teams

We are group 12 in the spreadsheet.
The UNI's of the group member are as follow:
1. ak4479
2. ht2459
3. sy2657
4. jc5066
5. yg2655

2. Read Chapter 1 of Statistical Sleuth

Completed

3. Data Questions for Display 1.3

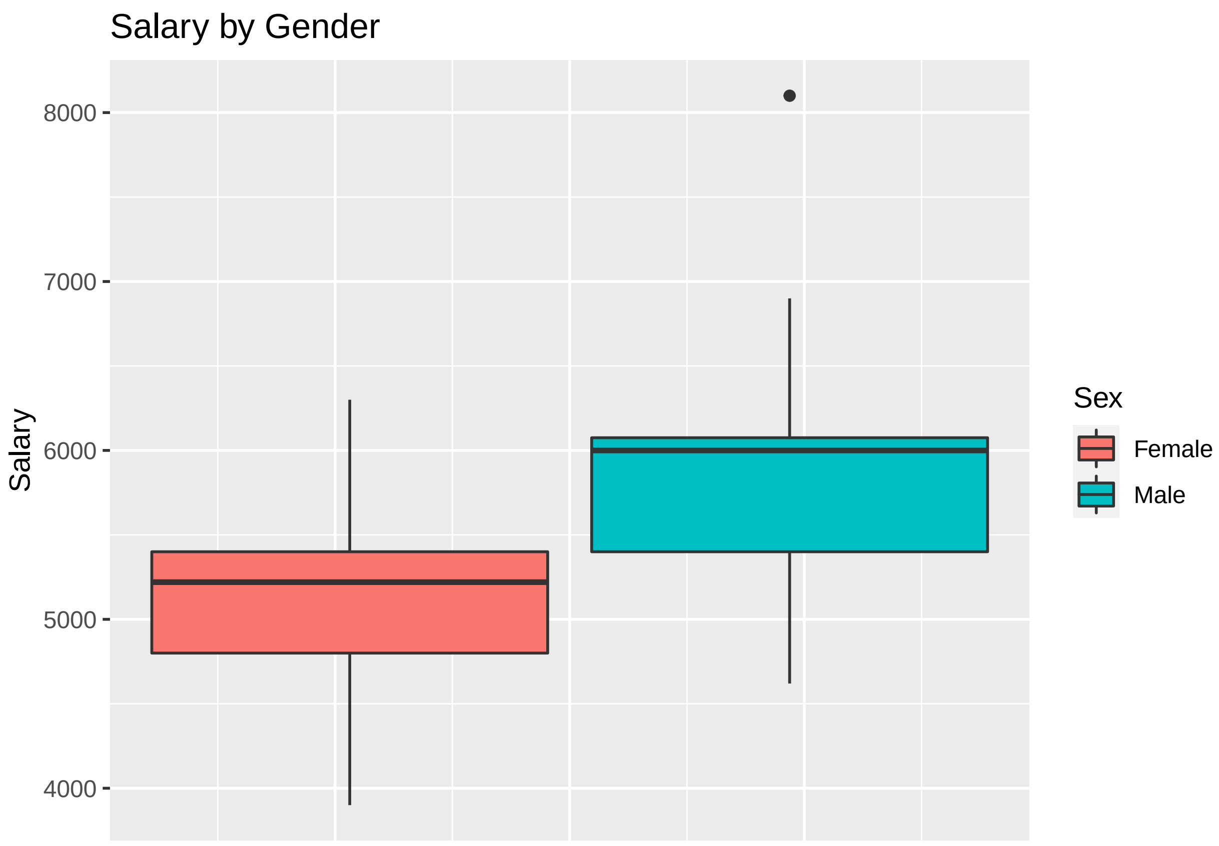
i) Determine whether there are outliers in the combined data, using boxplots or other suitable methods.

Using simple boxplots we can see that in the male category there is one salary of 8100 that is above the Upper Limit, which can be considered an outlier.

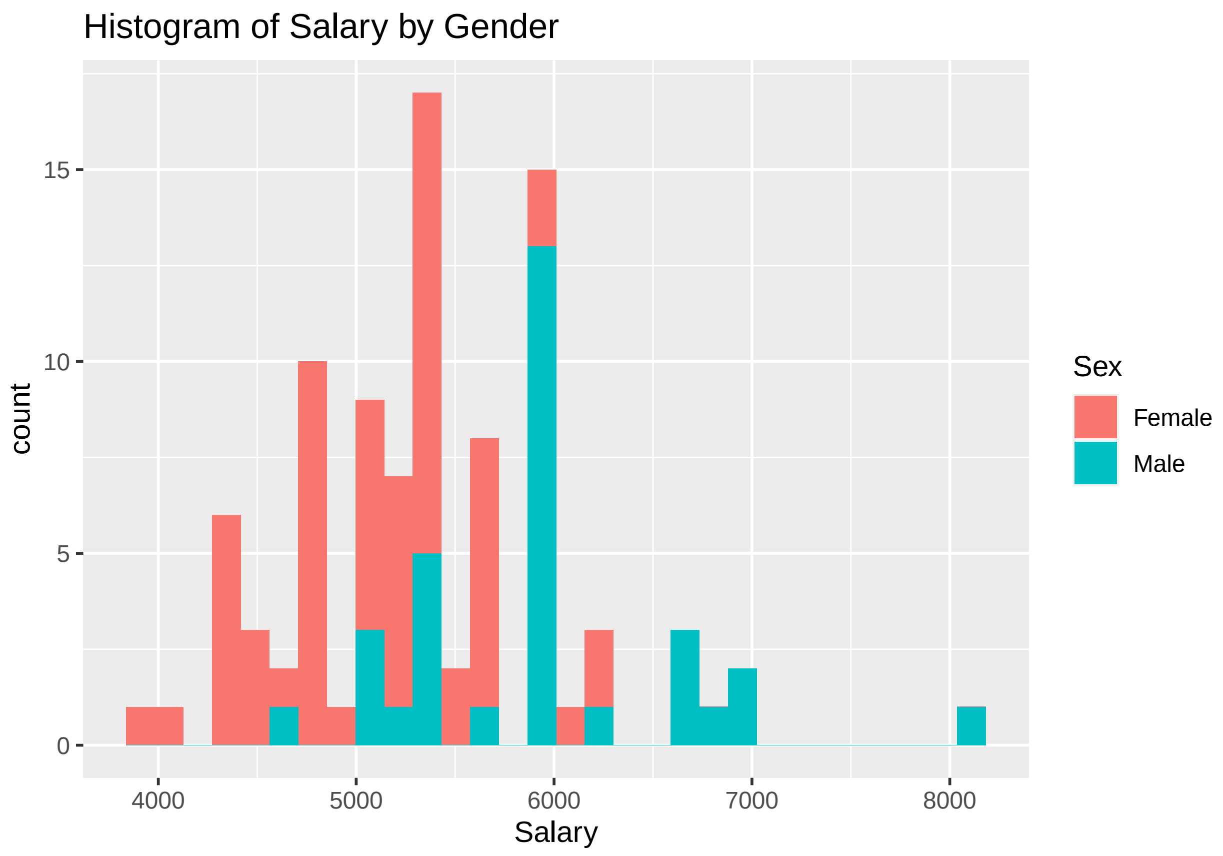
```r
library(Sleuth3)
library(ggplot2)
library(dplyr)
```

```
##
## 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
df <- case0102

p <- ggplot(data = df, aes(x = Salary,
                           fill = Sex))
p +geom_boxplot() + coord_flip() + theme(axis.title.x=element_blank(),
                                         axis.text.x=element_blank(),
                                         axis.ticks.x=element_blank()) +
  labs(title = "Salary by Gender")
```



```r
p + geom_histogram(bins = 30) + labs(title = "Histogram of Salary by Gender")
```



ii) Perform separate EDA, and compute the sample coefficient of variation and median for Salary in each group (i.e., Males and Females)

Below is a table of Standard Deviation, Mean, Coefficient of Variation, and Median of both genders in the data

```
sample_coef <- df %>% group_by(Sex) %>% summarise(SD  = sd(Salary),
                                                  mu = mean(Salary),
                                                  Coeff_var = SD / mu,
                                                  Median = median(Salary))
sample_coef
```

```
## # A tibble: 2 x 5
##   Sex       SD    mu Coeff_var Median
##   <fct>  <dbl> <dbl>     <dbl>  <dbl>
## 1 Female  540. 5139.     0.105   5220
## 2 Male    691. 5957.     0.116   6000
```

iii) For each of the estimates computed in (ii) above, determine the bias and variance using each of the following methods:

A. Jackknife

B. Bootstrap

```
library(purrr)
Num_Samples <- 10
```

```
K <- 1000
n <- nrow(df)

l <- list()
for(i in 1:Num_Samples){
  boot <- sample(1:n, size = K, replace = TRUE)
  boot_df <- df[boot,]
  boot_coeff <- boot_df %>% group_by(Sex) %>% summarise(SD  = sd(Salary),
                                                  mu = mean(Salary),
                                                  Coeff_var = SD / mu,
                                                  Median = median(Salary))
  l[[i]] <- boot_coeff
}
boot_coeff_combined <- bind_rows(l, .id = "Indicator") %>% select(-"Indicator") %>% group_by(Sex) %>%
  summarise(Coeff_var = mean(Coeff_var),
            Median = mean(Median))
boot_coeff_combined
```

```
## # A tibble: 2 x 3
##   Sex    Coeff_var Median
##   <fct>      <dbl>  <dbl>
## 1 Female     0.104   5214
## 2 Male       0.112   6000
```

```
(boot_coeff_combined %>% select(Coeff_var, Median)) - (sample_coef %>% select(Coeff_var, Median))
```

```
##      Coeff_var Median
## 1 -0.001533042     -6
## 2 -0.003533157      0
```