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

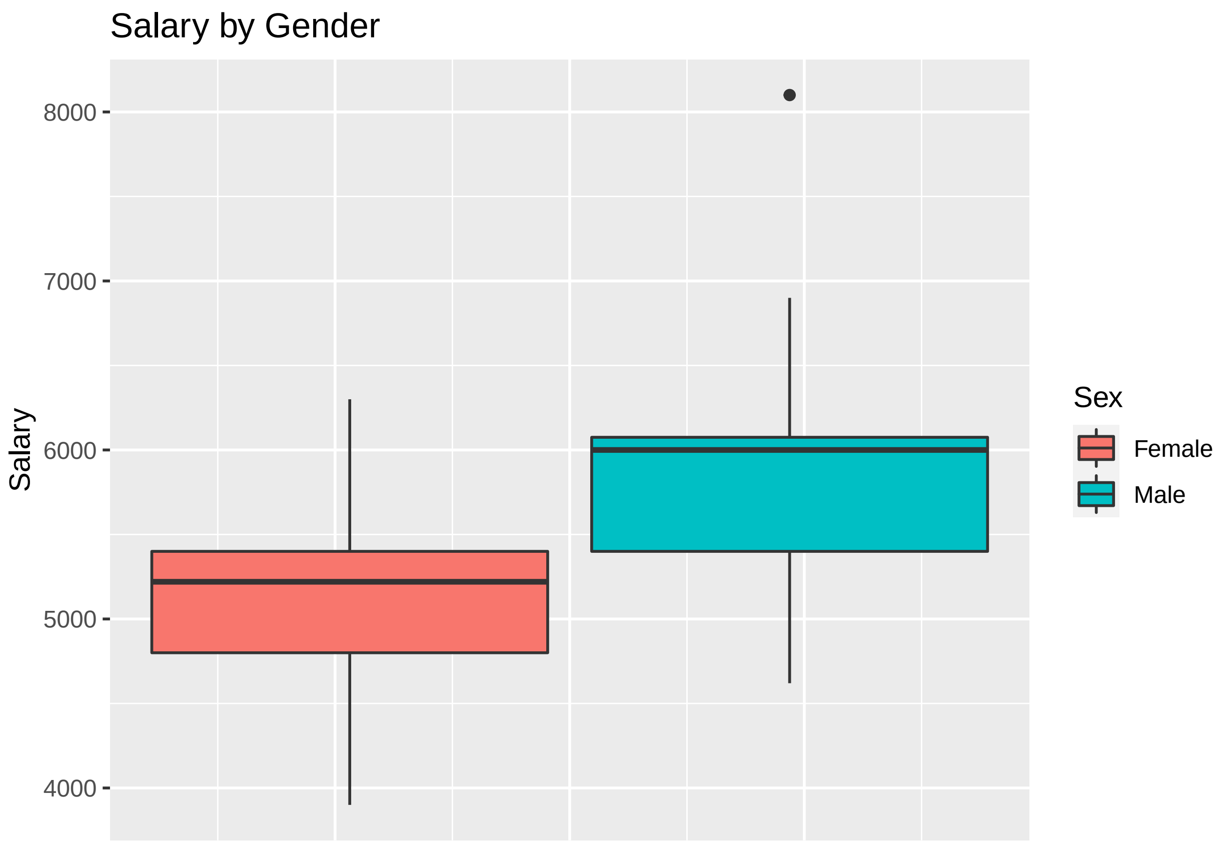
i) Determine whether there are outliers in the combined data, using boxplots or other suitable methods.

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.

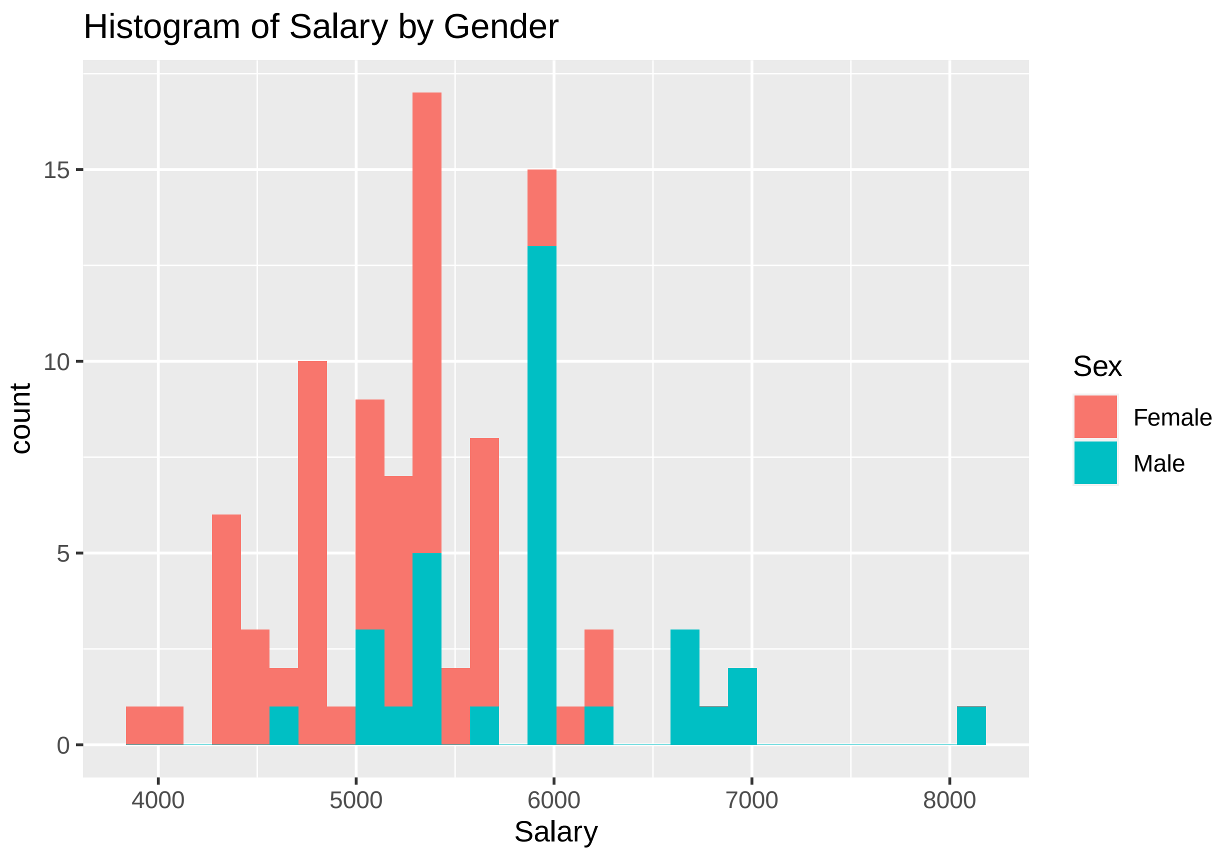
```
###############################################
#
# Creates boxplot/histogram to identify outliers
#
###############################################

# Load Relevant Libraries/Data
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

# Initialize Boxplot
p <- ggplot(data = df, aes(x = Salary,
                           fill = Sex))

# Plot Boxplot
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
# Plot Histogram
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

```
##################################################
#
# Calculates CV & Median of Data
#
##################################################

# Calculate Coefficient Variation and Median of original 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

Computed Jackknife estimate of Median and Coefficient Variation and their respective Variance/Bias. The final table outputs the Variance/Bias of the CV followed by the Median.

```
#######################################################
#
# Jackknife Simulation for Variance/Bias of CV & Median
#
#######################################################

# Initialize Parameters
K <- 1000
n <- nrow(df)
l <- list()

# Generate Jackknife Samples
for(i in 1:n){
  temp <- df[-i,]
  temp <- temp %>% group_by(Sex) %>% summarise(SD  = sd(Salary),
                                                  mu = mean(Salary),
                                                  Coeff_var = SD / mu,
                                                  Median = median(Salary))
  l[[i]] <- temp
}

# Calculate Resulting Coefficient Variation & Median variance/Bias
Jack_coeff_combined <- bind_rows(l, .id = "Indicator") %>% select(-"Indicator")

Jack_coeff_combined %>% group_by(Sex) %>%
  summarise(Var_Coeff_var = var(Coeff_var),
            Bias_Coeff_var = mean(Coeff_var) -
              unlist((sample_coef %>% filter(Sex == "Female") %>% select(Coeff_var))),
            Var_Median = var(Median),
            Bias_Median = mean(Median) -
              unlist((sample_coef %>% filter(Sex == "Female") %>% select(Median))))
```

```
## # A tibble: 2 x 5
##   Sex    Var_Coeff_var Bias_Coeff_var Var_Median Bias_Median
##   <fct>          <dbl>          <dbl>      <dbl>       <dbl>
## 1 Female   0.000000917    -0.00000386       202.         -10
## 2 Male     0.00000460      0.0109             0          780
```

B. Bootstrap

Computed Bootstrap estimate of Median and Coefficient Variation and their respective Variance/Bias. The final table outputs the Variance/Bias of the CV followed by the Median.

```
#######################################################
#
# Bootstrap Simulation for Variance/Bias of CV & Median
#
#######################################################
```

```r
# Initialize Parameters and Load Library
library(purrr)
Num_Samples <- 100
K <- 1000
n <- nrow(df)
l <- list()

# Generate Bootstrap samples
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
}

# Calculate Resulting Coefficient Variation & Median variance/Bias
boot_coeff_combined <- bind_rows(l, .id = "Indicator") %>% select(-"Indicator")

boot_coeff_combined %>% group_by(Sex) %>%
  summarise(Var_Coeff_var = var(Coeff_var),
            Bias_Coeff_var = mean(Coeff_var) -
              unlist((sample_coef %>% filter(Sex == "Female") %>% select(Coeff_var))),
            Var_Median = var(Median),
            Bias_Median = mean(Median) -
              unlist((sample_coef %>% filter(Sex == "Female") %>% select(Median))))
```

```
## # A tibble: 2 x 5
##   Sex    Var_Coeff_var Bias_Coeff_var Var_Median Bias_Median
##   <fct>          <dbl>          <dbl>      <dbl>       <dbl>
## 1 Female    0.00000621      -0.000557      2168.       -19.8
## 2 Male      0.0000280        0.00870          0        780
```