HW3_Salem

Mohamed Salem

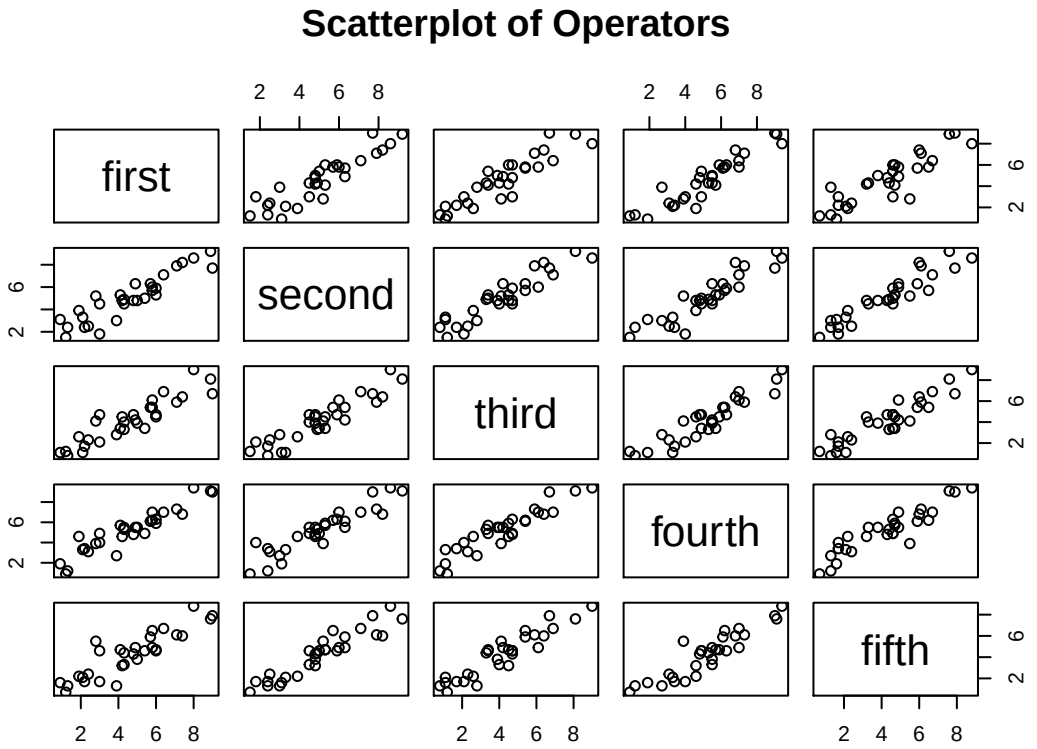
September 16, 2019

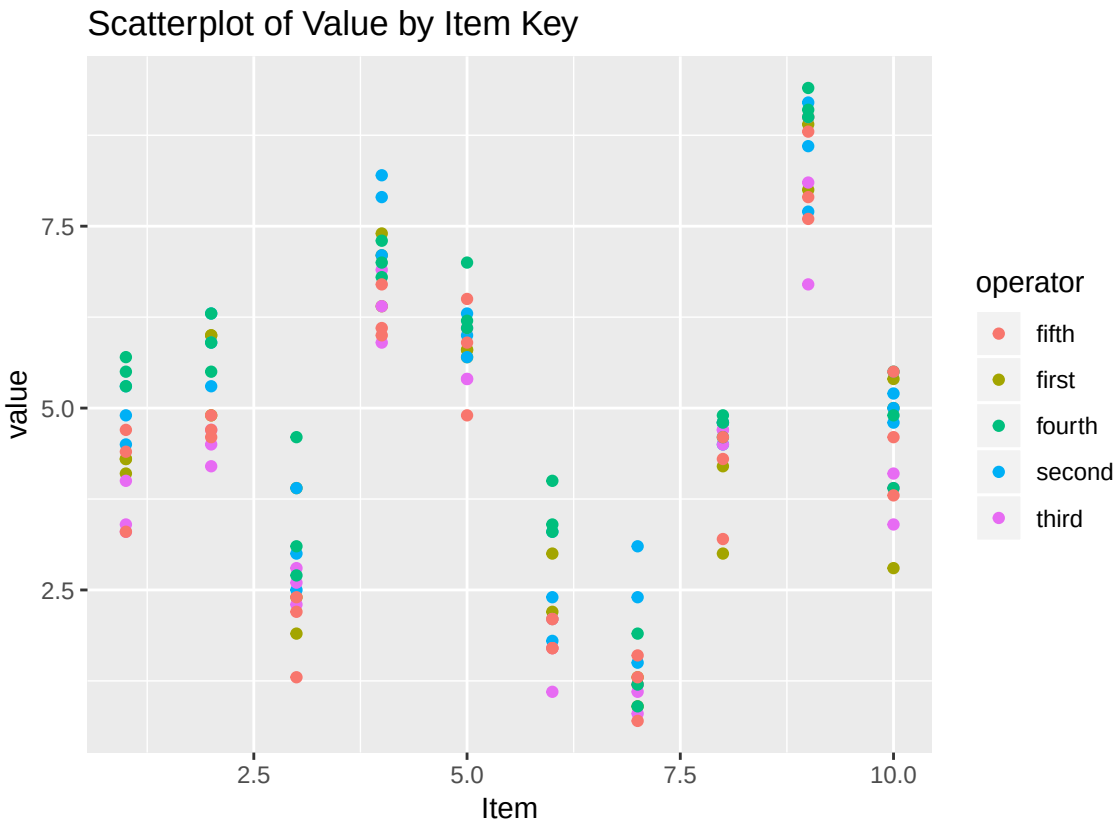
Problem 3 - Data Wrangling

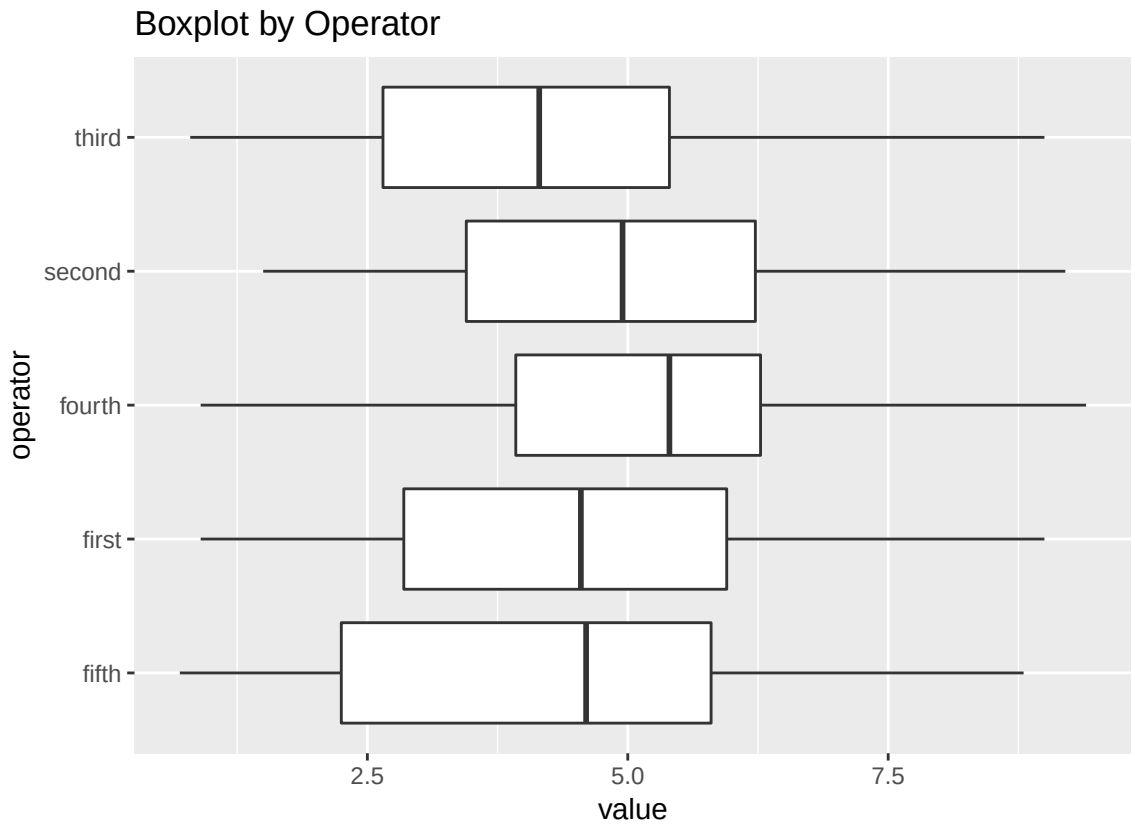
```
# The next line specifies the url which the data will be imported from
url1 <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"

# This creates a vector to hold the names available in the header of the data
name_placeholder <- c("Item", "first", "second", "third", "fourth", "fifth")

# This reads data from the source table after skipping the first line which is
# the header
Sensory_data_1 <- read.table(url1, fill = T, header = T, skip = 1, na.string = c("",
    "null", "NaN"), stringsAsFactors = FALSE, sep = " ")

# This applies our previously stored names to the imported dataframe
names(Sensory_data_1) <- name_placeholder

# The following uses dplyr and tidyr to clean the data
Sensory_data_1 <- as_tibble(Sensory_data_1)
Sensory_data_p1 <- Sensory_data_1 %>% filter(is.na(fifth))
Sensory_data_p2 <- Sensory_data_1 %>% filter(!is.na(fifth))
Sensory_data_p1 <- Sensory_data_p1 %>% mutate(adjitem = ceiling(row_number()/2)) %>%
    select(adjitem, Item, first, second, third, fourth)
names(Sensory_data_p1) <- name_placeholder
Sensory_data_1 <- bind_rows(Sensory_data_p2, Sensory_data_p1) %>% arrange(Item)

Sensory_data_1 <- bind_rows(Sensory_data_p2, Sensory_data_p1) %>% arrange(Item)

Sensory_data_2 <- Sensory_data_1 %>% gather(Item) %>% mutate(operator = Item) %>%
    mutate(Item = ceiling(row_number()/3)) %>% mutate(Item = Item - 10 * floor(Item/10)) %>%
    mutate(Item = Item + (Item == 0) * 10)
```

```
# Finally we create a summary table of our data grouped by operator...
summarise(group_by(Sensory_data_2, operator), count = n(), mean(value))
```

```
## # A tibble: 5 x 3
##   operator count `mean(value)`
##   <chr>    <int>         <dbl>
## 1 fifth       30          4.27
## 2 first       30          4.59
## 3 fourth      30          5.19
## 4 second      30          5.06
## 5 third       30          4.17
```

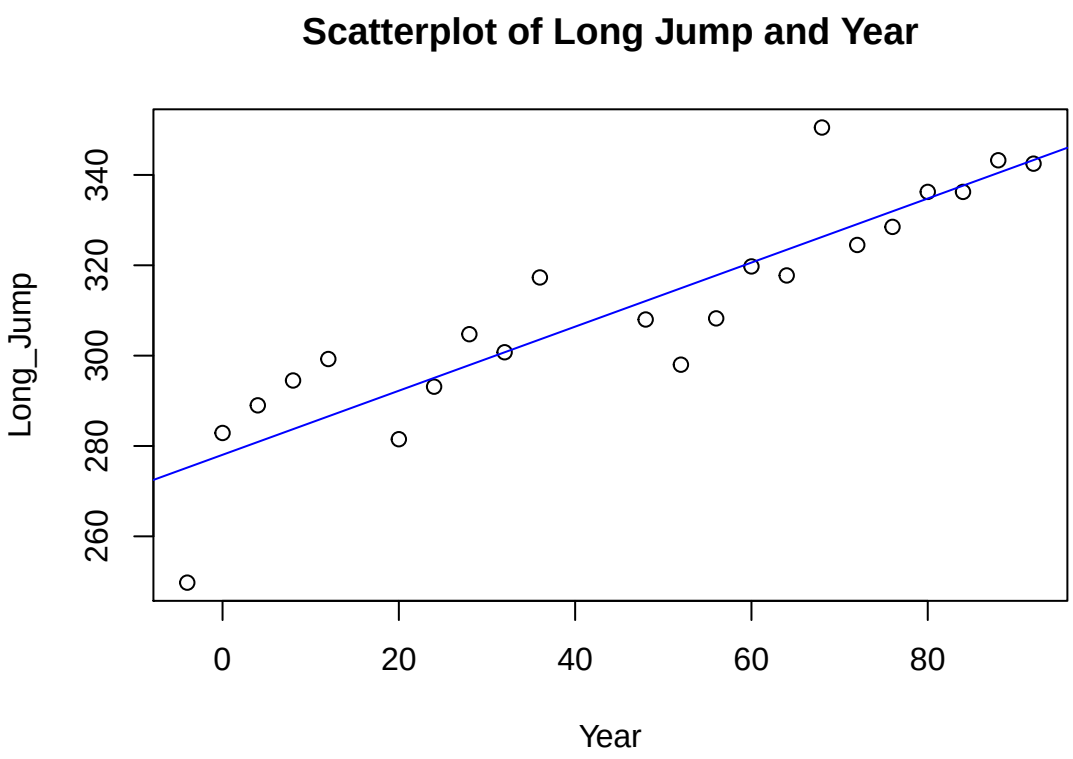

Scatterplot of Operators
first
second
third
fourth
fifth
2
4
6
8


Scatterplot of Value by Item Key
value
Item
2.5
5.0
7.5
10.0
operator
fifth
first
fourth
second
third




```
# The next line specifies the url which the data will be imported from
url2 <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"

# This reads data from the source table
Gold_medal_1 <- read.table(url2, fill = T, header = T, na.string = c("", "null",
    "NaN"), stringsAsFactors = FALSE, sep = " ")

# This creates a vector to hold the names we want to use as headers of the data
name_placeholder2 <- c("Year", "Long_Jump")

# The following uses dplyr and tidyr to clean the data
Gold_medal_p1 <- Gold_medal_1 %>% select(Year = Year, Long_Jump = Long)
Gold_medal_p2 <- Gold_medal_1 %>% select(Year = Jump, Long_Jump = Year.1)
Gold_medal_p3 <- Gold_medal_1 %>% select(Year = Long.1, Long_Jump = Jump.1)
Gold_medal_p4 <- Gold_medal_1 %>% select(Year = Year.2, Long_Jump = Long.2)
Gold_medal_1 <- bind_rows(Gold_medal_p1, Gold_medal_p2, Gold_medal_p3, Gold_medal_p4) %>%
    filter(!(is.na(Year)))

# Finally we create a summary table of our data ...
summary(Gold_medal_1$Long_Jump)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   249.8   295.4   308.1   310.3   327.5   350.5
```



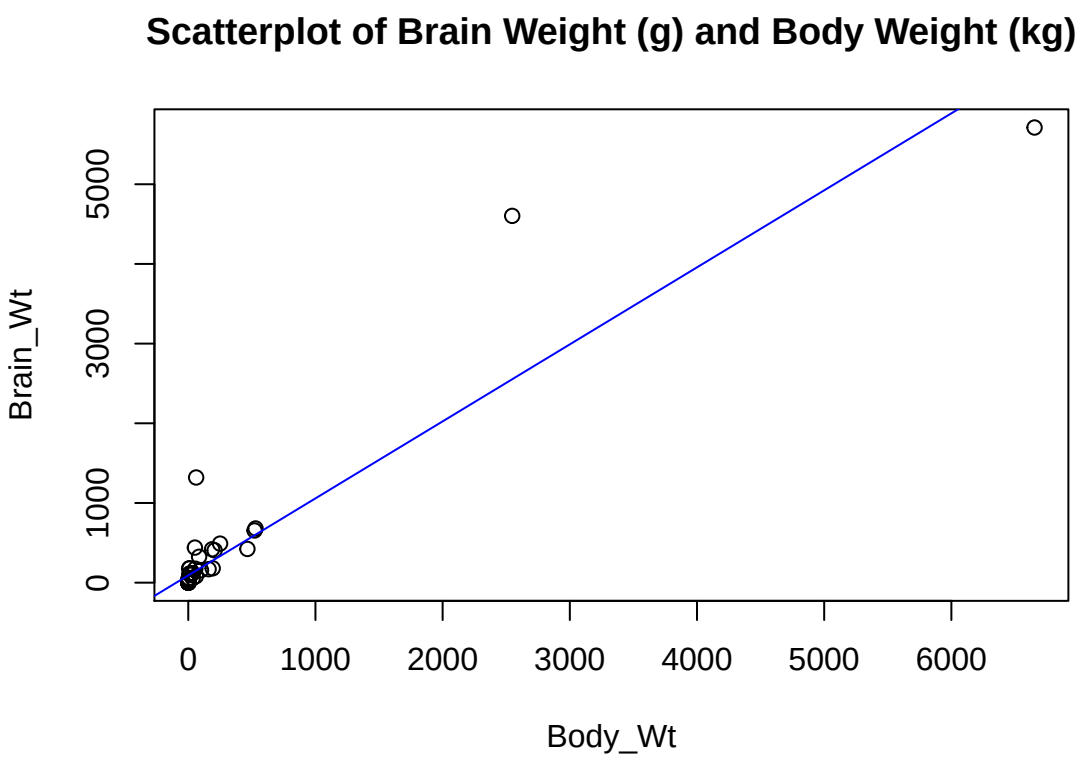

```
# The next line specifies the url which the data will be imported from
url3 <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat"

# This reads data from the source table
Brain_weight_1 <- read.table(url3, fill = T, header = T, na.string = c("", "null",
    "NaN"), stringsAsFactors = FALSE, sep = " ")

# This creates a vector to hold the names we want to use as headers of the data
name_placeholder3 <- c("Body_Wt_kg", "Brain_Wt_g")

# The following uses dplyr and tidyr to clean the data
Brain_weight_p1 <- Brain_weight_1 %>% select(Body_Wt = Body, Brain_Wt = Wt)
Brain_weight_p2 <- Brain_weight_1 %>% select(Body_Wt = Brain, Brain_Wt = Wt.1)
Brain_weight_p3 <- Brain_weight_1 %>% select(Body_Wt = Body.1, Brain_Wt = Wt.2)
Brain_weight_1 <- bind_rows(Brain_weight_p1, Brain_weight_p2, Brain_weight_p3) %>%
    filter(!(is.na(Brain_Wt)) & !(is.na(Body_Wt)))
```

```
# Finally we create a summary table of our data ...
summary(Brain_weight_1$Body_Wt)
```

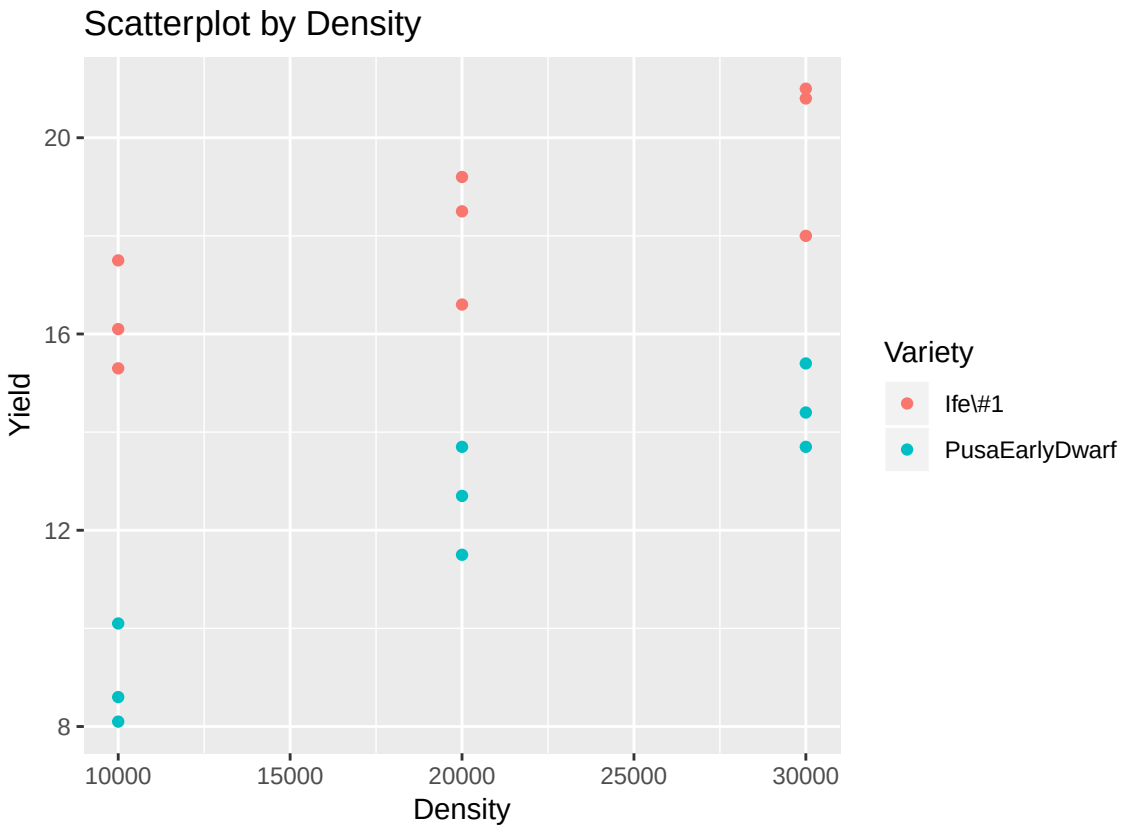
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.005    0.600    3.342  198.790   48.203 6654.000
```

```
summary(Brain_weight_1$Brain_Wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.10    4.25   17.25  283.13  166.00 5712.00
```




```
# The next line specifies the url which the data will be imported from
url4 <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat"

# This reads data from the source table
firstrow1 <- read.table(url4, skip = 2, nrow = 1, quote = "", comment.char = "",
    stringsAsFactors = T, sep = c(" ", ";"))
secondrow1 <- read.table(url4, skip = 3, nrow = 1, stringsAsFactors = FALSE,
    sep = c(" ", ";"))

# The following uses dplyr and tidyr to clean the data
Triplicate_1 <- bind_rows(firstrow1, secondrow1) %>% separate(V4, c("one",
    "two", "three"), sep = ",") %>% separate(V8, c("four", "five", "six"),
    sep = ",") %>% separate(V11, c("seven", "eight", "nine"), sep = ",") %>%
    separate(V12, c("ten", "eleven", "twelve"), sep = ",") %>% separate(V15,
    c("thirteen", "fourteen", "fifteen"), sep = ",") %>% separate(V18,
    c("sixteen", "seventeen", "eighteen"), sep = ",") %>% select(1, one,
    two, three, four, five, six, seven, eight, nine, ten, eleven, twelve,
    thirteen, fourteen, fifteen, sixteen, seventeen, eighteen) %>% gather(key = Plant_type,
    value = Value, 2:19) %>% select(V1, Value) %>% filter(!is.na(Value)) %>%
    mutate(Density = (floor((row_number() - 1)/3) + 1) * 10000) %>% mutate(Density = Density -
    (Density == 40000) * 30000) %>% mutate(Density = Density - (Density ==
    50000) * 30000) %>% mutate(Density = Density - (Density == 60000) *
    30000) %>% mutate(Value = as.numeric(Value))

# This names the headers of our data
names(Triplicate_1) <- c("Variety", "Yield", "Density")
```


Scatterplot by Density
Yield
20
16
12
8
10000
15000
20000
25000
30000
Density
Variety
Ife\#1
PusaEarlyDwarf