# Data Analysis Playground

Khin Mon



```
# loading all required packages
if(!require(knitr)){
    install.packages("knitr")
    library(knitr)
}

if(!require(kableExtra)){
    install.packages("kableExtra", repos = "https://www.stats.bris.ac.uk/R/")
    library(kableExtra)
}

if(!require(psych)){
    install.packages("psych")
    library(psych)
}

if(!require(ggplot2)){
    install.packages("ggplot2")
    library(ggplot2)
}

# for melt
if(!require(reshape2)){
    install.packages("reshape2")
    library(reshape2)
}
```

# Introduction

The provided dataset has 2000 observations with one categorical variable and 16 numeric variables (10 continuous and 6 discrete variables).

# Data Analysis

The first step is we loaded the dataset and we will do the basic data analysis step by step to explore the provided dataset.

```
# loading the dataset
loaded_data <- read.csv("dataset/dataset.csv", stringsAsFactors = TRUE)
#Loaded_data

# splitting data only and label
loaded_df <-loaded_data[,-ncol(loaded_data)]
loaded_label <-loaded_data[,ncol(loaded_data)] # Group label
```

# Descriptive Analysis

```
kbl(head(loaded_data), caption="Loaded Dataset with some observations") %>% kable_classic(full_w
idth = F, html_font = "Cambria")
```

Loaded Dataset with some observations

| Item1 | Item2 | Item3 | Item4 | Item5 | Item6 | Item7 | Item8 | Item9 | Item10 | b1 | b2 | b3 | b4 | b5 | b6 | Group |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 14.5 | 2.0 | 11.0 | 5.0 | 5.0 | 8.0 | 14.5 | 11.0 | 5.0 | 11.0 | 0 | 7 | 3 | 9 | 13 | 1 | A |
| 14.5 | 4.5 | 8.5 | 4.5 | 8.5 | 8.5 | 14.5 | 12.0 | 2.0 | 8.5 | 0 | 6 | 3 | 11 | 13 | 1 | A |
| 11.5 | 7.5 | 14.5 | 1.0 | 4.0 | 11.5 | 14.5 | 7.5 | 7.5 | 7.5 | 0 | 5 | 3 | 10 | 13 | 2 | A |
| 14.5 | 7.0 | 11.5 | 1.0 | 7.0 | 7.0 | 14.5 | 11.5 | 7.0 | 7.0 | 0 | 4 | 3 | 10 | 13 | 2 | A |
| 14.0 | 2.0 | 8.0 | 0.0 | 8.0 | 5.0 | 14.0 | 14.0 | 8.0 | 11.0 | 1 | 6 | 4 | 10 | 12 | 3 | A |
| 11.5 | 8.0 | 14.5 | 4.5 | 2.0 | 11.5 | 14.5 | 8.0 | 4.5 | 8.0 | 0 | 6 | 3 | 10 | 13 | 1 | A |

In the loaded dataset, there are 2000 observations with 17 variables which consists of:

- 1 label,
- 10 continuous and
- 6 discrete variables.

```
# summarize data
summary_data <- describe(loaded_data[,-ncol(loaded_data)], IQR=TRUE , skew=TRUE)
summarydata <- summary_data[,c(3,4,8,5,9,13, 14)]
kbl(summarydata, caption = "Summary Statistic of the dataset") %>% kable_classic(full_width = F,
html_font = "Cambria")
```

Summary Statistic of the dataset

| | mean | sd | min | median | max | se | IQR |
|---|---|---|---|---|---|---|---|
| Item1 | 11.98850 | 1.8693221 | 5.0 | 12.0 | 15.0 | 0.0417993 | 2.000 |
| Item2 | 3.94525 | 2.4440467 | 0.0 | 4.0 | 12.0 | 0.0546505 | 3.500 |
| Item3 | 10.19300 | 2.3711415 | 2.5 | 10.5 | 15.0 | 0.0530203 | 3.000 |
| Item4 | 1.49850 | 1.7052717 | 0.0 | 1.0 | 8.5 | 0.0381310 | 2.000 |
| Item5 | 4.00025 | 2.4059342 | 0.0 | 4.0 | 11.5 | 0.0537983 | 3.500 |
| Item6 | 6.97800 | 2.7690431 | 0.5 | 7.0 | 14.5 | 0.0619177 | 3.625 |
| Item7 | 13.92175 | 1.2983046 | 5.5 | 14.5 | 15.0 | 0.0290310 | 1.500 |
| Item8 | 9.24850 | 2.3108803 | 1.0 | 9.0 | 14.5 | 0.0516729 | 3.000 |
| Item9 | 4.97575 | 2.2490701 | 0.0 | 5.0 | 11.0 | 0.0502907 | 3.500 |
| Item10 | 6.02150 | 2.4270833 | 0.0 | 6.0 | 14.0 | 0.0542712 | 3.500 |
| b1 | 1.10150 | 1.1661379 | 0.0 | 1.0 | 6.0 | 0.0260756 | 2.000 |

| | mean | sd | min | median | max | se | IQR |
|---|---|---|---|---|---|---|---|
| b2 | 9.37700 | 2.3342226 | 4.0 | 10.0 | 13.0 | 0.0521948 | 2.000 |
| b3 | 6.76800 | 2.4525871 | 2.0 | 7.0 | 12.0 | 0.0548415 | 3.000 |
| b4 | 11.88850 | 1.6883403 | 7.0 | 12.0 | 14.0 | 0.0377524 | 2.000 |
| b5 | 14.15700 | 0.8704768 | 12.0 | 14.0 | 15.0 | 0.0194645 | 1.000 |
| b6 | 3.93700 | 2.3276039 | 1.0 | 3.0 | 10.0 | 0.0520468 | 4.000 |

According to the above summary table, there are no extreme values between minimum and maximum of each variable. So, even if there might be the outliers in the variables, those might not be so extreme to get removed from the dataset. Low standard deviation means observations are close to the mean.

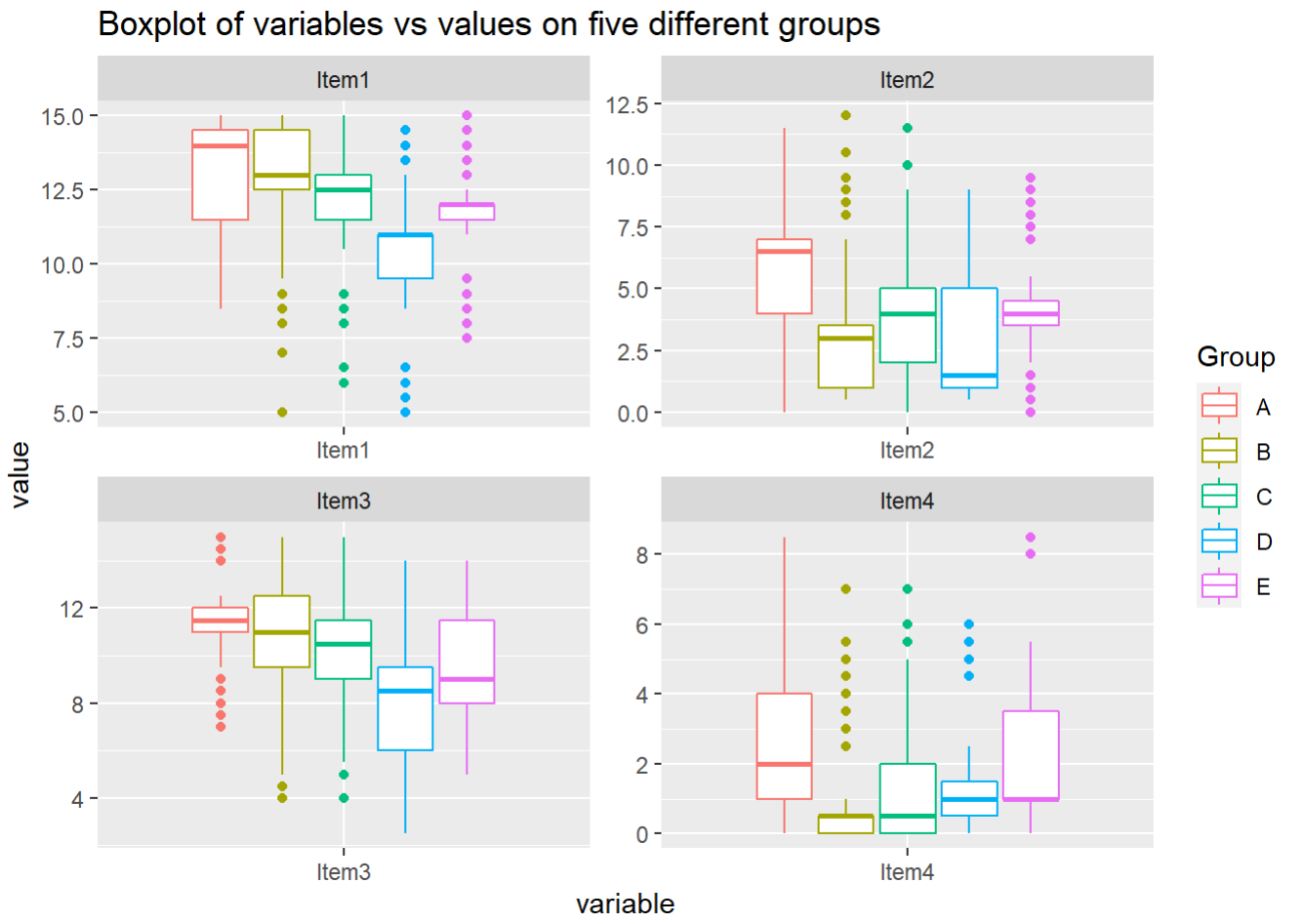
# Exploratory Data Analysis

## Missing Values

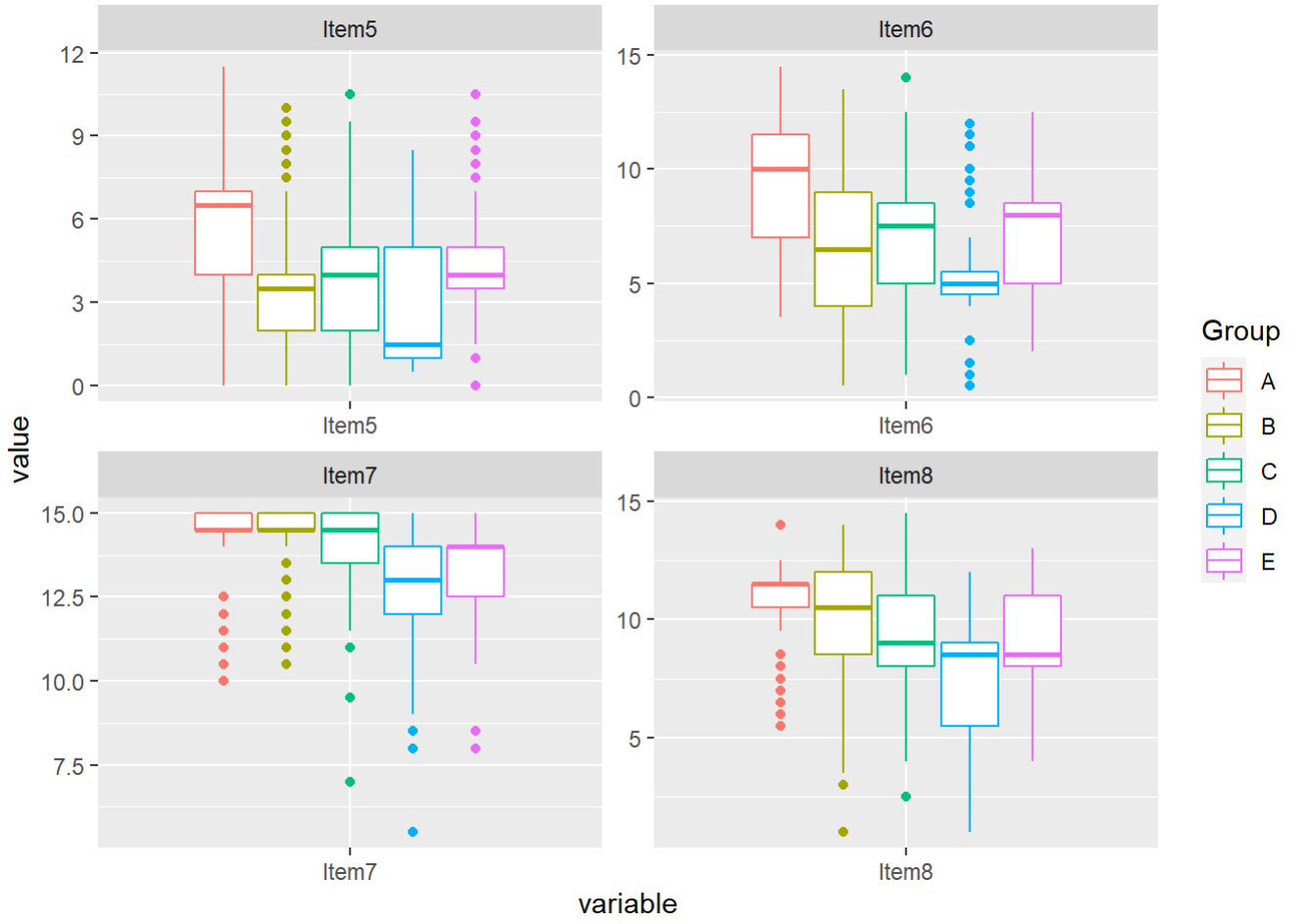
There are 0 missing values in this original dataset.

## Analysing outliers and data distributions

```
# boxplots, item 1 to 4
ggplot(melt(loaded_data[,c(1:4, 17)], id.var = "Group"), aes(x=variable, y=value, color = Grou
p)) +
  geom_boxplot() +
  facet_wrap( ~variable, scales="free") +
  labs(title = "Boxplot of variables vs values on five different groups")
```




```
# boxplots, item 5 to 8
ggplot(melt(loaded_data[,c(5:8, 17)], id.var = "Group"), aes(x=variable, y=value, color = Grou
p)) +
  geom_boxplot() +
  facet_wrap( ~variable, scales="free")
```

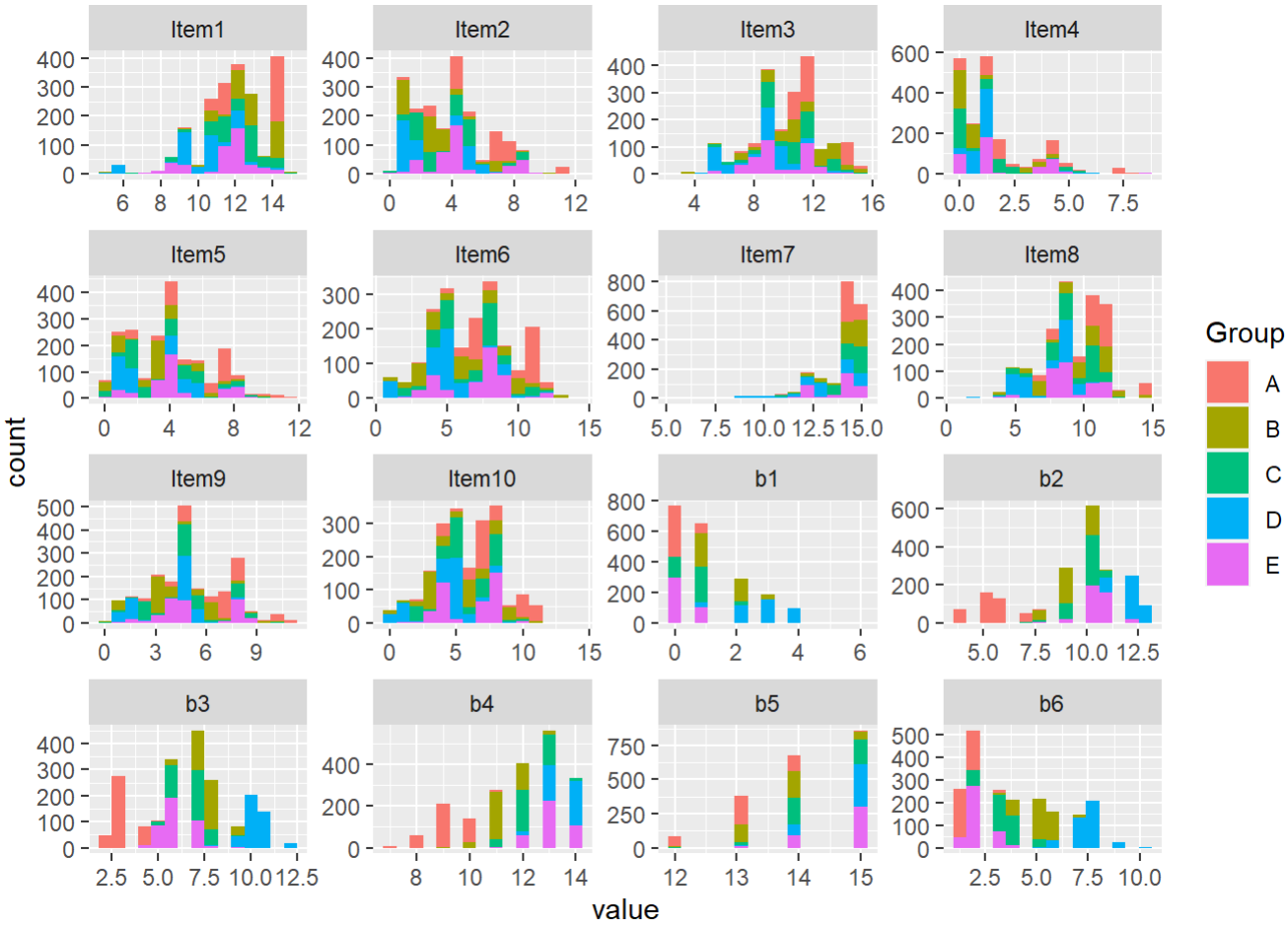



```
# boxplots, item 9 to 12 [pls try yourself]

# boxplots, item 13 to 16 [pls try yourself]
```

From the above boxplots, we could see that almost all variables include the outliers. Some are left or right skewed distributions and apparently, there is no variable which follows normal distribution(bell-shaped).

```
ggplot(melt(loaded_data, id.var = "Group"), aes(value)) +
  geom_histogram(bins=15, aes(fill=Group)) +
  facet_wrap(~variable, scales = "free")
```




[to be continued] …