PROJECT REPORT - THE OUTLIERS

“Analysis of the relationship between user satisfaction and effectiveness using drug type as a factor”

Team Members

Balakumar Rahul

Geetaharichandana Vemuri

Sreekar Adusumilli

Upamanyu Mondal

Venkataramana Pittala

Department of Health Informatics,

IU Luddy School of Informatics, Computing, and Engineering,

IUPUI.

Timothy Gruenhagen

May 04, 2023

INTRODUCTION

Over the past years, there has been a growing interest in how satisfied patients are with their medicine or medical device. This reflects the growing consumerization of healthcare and the desire of pharmaceutical and device companies to hear from customers about their goods. It is important to distinguish drug satisfaction from other types of satisfaction since it is more specific (Shikiar et al., 2004).

Several studies have explored the prevalence and patterns of prescription and OTC drug use in different populations. For instance, a study by Qato et al. (2016) found that polypharmacy (multiple prescription drugs) is more common among older adults and those with multiple chronic conditions. Another study by Zafeiri et al. (2021) examined OTC drug use among pregnant women and found that a significant proportion of them use non-recommended drugs, such as aspirin and ibuprofen, which can have adverse effects on fetal development.

Our initiative aims to examine the efficacy and user satisfaction of various drug types used to treat 37 prevalent medical illnesses. User satisfaction and effectiveness are crucial elements that might affect a patient's adherence to medication and general health results, as was previously discussed. Therefore, our project will assess the efficacy and satisfaction level of both over-the-counter (OTC) and prescription (Rx) medications, by assessing the user ratings received from patients. In the long run, this information can aid clinical judgment and enhance patient outcomes.

DATA VARIABLES

The DRUG PERFORMANCE EVALUATION DATASET was obtained from Kaggle. It contains drug performance characteristics for different types of drugs used to treat 37 common medical conditions. It has data on user satisfaction and effectiveness collected from patients. The dataset contains drug information, including medical conditions, names, indications, types, reviews, effectiveness, ease of use, and satisfaction.

METHODS

Data Cleaning and Organization

We started our analysis by first coding the “Types” column in R. “RX” i.e. prescription drug, was coded as 0, “OTC” was coded as 1, and “OTC/RX” was coded as 2. Next, checked for any duplicates or blank values, but we did not find any.

Exploratory Data Analysis

Scatter Plot

Next, we proceeded with the exploratory data analysis. We plotted a scatter plot between the two continuous variables i.e., Satisfaction and Effectiveness. We saw that most of the values are between 2-4. There were few values with 1 and 5.

Boxplot

We further checked for the distribution of the Effectiveness and the Satisfaction of the drug using a Boxplot.

We can see that drug effectiveness has 50 percent of its value between 3 to 4, whereas satisfaction has 50 percent of its value a little lower. This is seen in the median as well, which is higher for Effectiveness compared to Satisfaction.

Density plot

Density plots are generally used to visualize the distribution of continuous variables. The probability density function is shown on the y-axis and variable values on the x-axis. The peak on the x-axis indicates the most commonly found variable in the dataset, in our case that happened to be prescription(RX) for the drug types. For effectiveness and satisfaction, the most commonly found values are between 3 and 4 i.e., the user ratings on the scale of 5.

Checking Data Distribution

We then proceeded with histograms so as to check the distribution of effectiveness and satisfaction and a count plot for the types of drugs. We saw that most of the drugs are prescription drugs, followed by over-the-counter drugs, and only a few are RX/OTC sold both as prescription and over the counter as shown in Fig(1).

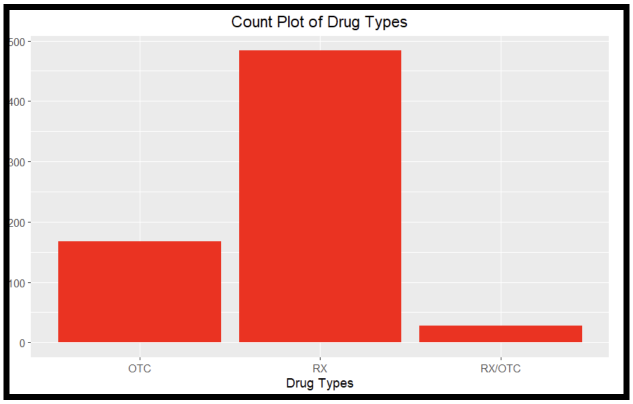


Fig 1

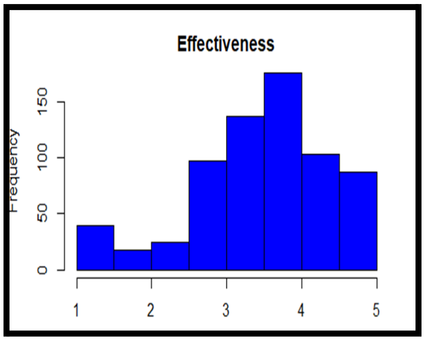


Fig 2

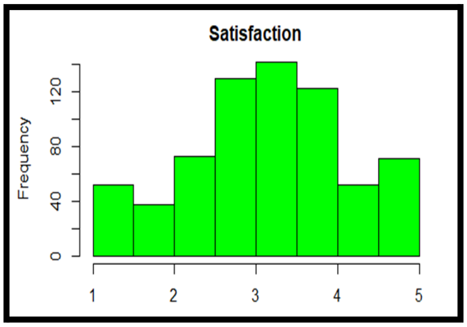


Fig 3

The histogram for effectiveness is shown as skewed whereas Satisfaction appears to be nearly normal as shown in Fig(2). We then checked for kurtosis and skewness and the results are provided in Fig(3). In this case, both the "Effective" and "Satisfaction" variables have negative skewness, suggesting a left-skewed distribution. The kurtosis of the "Effective" variable suggests a moderately peaked distribution with a somewhat heavier tail compared to a normal distribution. Whereas the kurtosis of the "Satisfaction" indicates a relatively flat distribution with a lighter tail compared to a normal distribution.

Shapiro-Wilk

To check for normality, we carried out Shapiro-Wilk Test. We carried out the test with the following two null hypotheses.

Null hypothesis(H0): the variable Effectiveness is normally distributed vs. H1; the variable is not normally distributed.

Null hypothesis(H0): the variable Satisfaction is normally distributed vs. H1; the variable is not normally distributed.

Since the p-value we deduced happened to be extremely small (less than 0.05), it suggests strong evidence to reject the null hypothesis of normality. Therefore, both the values "Effective" variable and “Satisfaction” is likely not normally distributed.

Kruskal-Wallis(Hypothesis Testing)

The Kruskal-Wallis test is a non-parametrical test that can be used for determining any significant differences between two or more groups. It was used as the data violated normality. We created a null and alternate hypothesis and checked for it using the results generated from the Kruskal-Wallis test.

Null Hypothesis: There are no differences in effectiveness and user satisfaction levels when

using a prescription, over-the-counter, or a combination of both drugs.

Alternate Hypothesis- There are differences in effectiveness and user satisfaction levels when

using a prescription, over-the-counter, or a combination of both drugs.

As the p-value was well below 0.05, we had sufficient enough evidence for us to reject the Null hypothesis and acknowledge that there were differences in effectiveness and satisfaction levels.

Relevance of the Statistical Models

We built a multinomial logistic regression model. The model tried to examine the relationship between one categorical response variable (status) with two continuous predictor variables (Effective and Satisfaction). The coefficients section shows the estimated coefficients for each individual predictor variable. It says how the log odds of each category of the response variable vary with an increase in the predictor variable. The standard errors column is used for showing the standard errors of the coefficients. The residual deviance and AIC are used to showcase how well the model fits the data and measures the trade-off between the goodness of fit and complexity. The lower the values, the better the fit. The AIC values yielded by the Generalized Linear Model(GLM) were on the higher side, hence, we decided to pursue ahead with the multinomial logistic regression model.

Limitations Of the Statistical Methods

Some common limitations of the statistical methods used in this study include:

1. The sample size may also not be representative of the general population, which might have limited the accuracy of these results.
2. Another limitation while using multinomial logistic regression is that it does not account for multiclass variables that have a natural ordering to them.

3. Other limitations of the statistical methods include data entry errors, coding errors, etc.

CONCLUSION

The use of prescription and over-the-counter (OTC) drugs has become increasingly common in modern society. According to the Centers for Disease Control and Prevention (CDC), approximately 70% of Americans take at least one prescription drug, and 30% take five or more. In addition, many people use OTC drugs to treat various conditions, including pain, fever, allergies, and gastrointestinal issues (CDC, 2019). The findings indicated that the majority of drugs in the dataset were prescription drugs, followed by OTC drugs, with only a small number being sold both as prescription and over the counter. The scatter plot and boxplot visualizations provided a glimpse into the relationship between satisfaction and effectiveness ratings, showing variations and trends within the dataset. Furthermore, the distribution analysis indicated that the effectiveness and satisfaction ratings had slightly left-skewed distributions. The Shapiro-Wilk test confirmed that both variables were not normally distributed. The Kruskal-Wallis test revealed significant differences in effectiveness and user satisfaction levels when comparing prescription, OTC, and combination drugs.

References:

Birkhäuer, J., Gaab, J., Kossowsky, J., Hasler, S., Krummenacher, P., Werner, C., & Gerger, H. (2017). Trust in the health care professional and health outcome: A meta-analysis. *PloS one*, *12*(2), e0170988. https://doi.org/10.1371/journal.pone.0170988

Centers for Disease Control and Prevention. (2019). FastStats - Therapeutic Drug Use. https://www.cdc.gov/nchs/fastats/drug-use-therapeutic.htm

Qato, D. M., Wilder, J., Schumm, L. P., Gillet, V., & Alexander, G. C. (2016). Changes in Prescription and Over-the-Counter Medication and Dietary Supplement Use Among Older Adults in the United States, 2005 vs 2011. *JAMA internal medicine*, *176*(4), 473–482. https://doi.org/10.1001/jamainternmed.2015.8581

Shikiar, R., & Rentz, A. M. (2004). Satisfaction with medication: an overview of conceptual, methodologic, and regulatory issues. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*, *7*(2), 204–215. https://doi.org/10.1111/j.1524-4733.2004.72252.x

Zafeiri, A., Mitchell, R. T., Hay, D. C., & Fowler, P. A. (2021). Over-the-counter analgesics during pregnancy: a comprehensive review of the global prevalence and offspring safety. Human reproduction update, 27(1), 67–95. https://doi.org/10.1093/humupd/dmaa042

APPENDIX SECTION

Project

2023-04-29

#Loading and reading the cleaned csv dataset from the specified directory

```
data<- read.csv("C:/Users/vibar/Downloads/Drug_clean.csv")

#Numerically encoding the Type of Drug column
data$Status <- ifelse(data$Type == "RX", 0,
               ifelse(data$Type %in% c("OTC"), 1, 2))

head(data)
```

```
##                   Condition                        Drug EaseOfUse Effective
## 1 Acute Bacterial Sinusitis                 Amoxicillin  3.852353  3.655882
## 2 Acute Bacterial Sinusitis Amoxicillin-Pot Clavulanate  3.470000  3.290000
## 3 Acute Bacterial Sinusitis Amoxicillin-Pot Clavulanate  3.121429  2.962857
## 4 Acute Bacterial Sinusitis                  Ampicillin  2.000000  3.000000
## 5 Acute Bacterial Sinusitis                  Ampicillin  3.250000  3.000000
## 6 Acute Bacterial Sinusitis           Ampicillin Sodium  3.000000  3.000000
##             Form Indication     Price   Reviews Satisfaction Type Status
## 1        Capsule   On Label  12.59000  86.29412     3.197647   RX      0
## 2 Liquid (Drink)  Off Label 287.37000  43.00000     2.590000   RX      0
## 3         Tablet   On Label  70.60857 267.28571     2.248571   RX      0
## 4        Capsule   On Label  12.59000   1.00000     1.000000   RX      0
## 5         Tablet   On Label 125.24000  15.00000     3.000000   RX      0
## 6         Tablet  Off Label 143.21500   1.00000     3.000000   RX      0
```

#Checking for Missing and Duplicate Values

```
#Missing Values
if (any(is.na(data))) {
  print("There are missing values in the dataset.")
} else {
  print("There are no missing values in the dataset.")
}
```

```
## [1] "There are no missing values in the dataset."
```

```
#For Duplicates
if (any(duplicated(data))) {
  print("There are duplicate rows in the dataset.")
} else {
  print("There are no duplicate rows in the dataset.")
}
```

```
## [1] "There are no duplicate rows in the dataset."
```

#Descriptive Statistics For both the continuous variables “Effective and”Satisfaction”

```r
# Calculate mean
mean1<- mean(data$Effective)
cat("The mean of the Effectiveness variable is:", mean1, "\n")
```

```
## The mean of the Effectiveness variable is: 3.52563
```

```r
mean2<- mean(data$Satisfaction)
cat("The mean of the Satisfaction variable is:", mean2, "\n")
```

```
## The mean of the Satisfaction variable is: 3.192844
```

```r
# Calculate median
median1<- median(data$Effective)
cat("The median of the Effectiveness variable is:", median1, "\n")
```

```
## The median of the Effectiveness variable is: 3.6
```

```r
median2<- median(data$Satisfaction)
cat("The median of the Satisfaction variable is:", median2, "\n")
```

```
## The median of the Satisfaction variable is: 3.2
```

```r
# Calculate range
range1<- range(data$Effective)
cat("The range of the Effectiveness variable is:", range1, "\n")
```

```
## The range of the Effectiveness variable is: 1 5
```

```r
range2<- range(data$Satisfaction)
cat("The range of the Satisfaction variable is:", range2, "\n")
```

```
## The range of the Satisfaction variable is: 1 5
```

```r
# Calculate standard deviation
sd1<- sd(data$Effective)
cat("The sd of the Effectiveness variable is:", sd1, "\n")
```

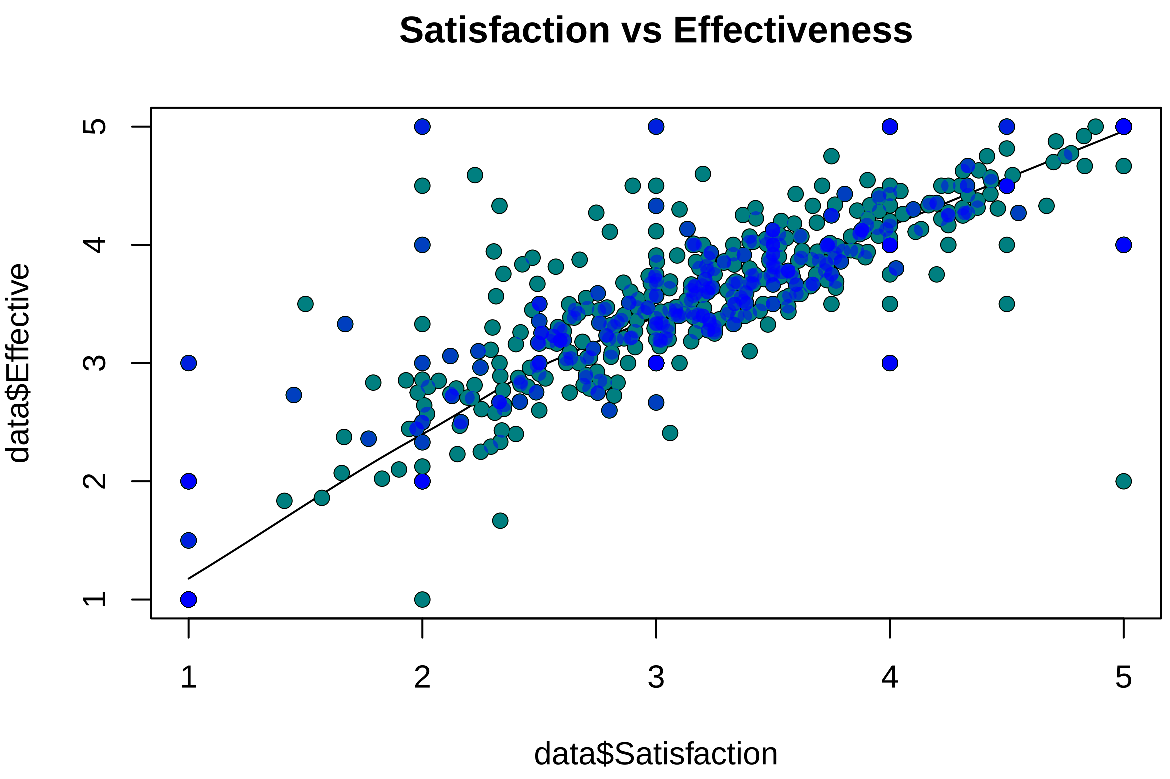
```
## The sd of the Effectiveness variable is: 0.9551967
```

```r
sd2<-sd(data$Satisfaction)
cat("The sd of the Satisfaction variable is:", sd2, "\n")
```

```
## The sd of the Satisfaction variable is: 1.030673
```

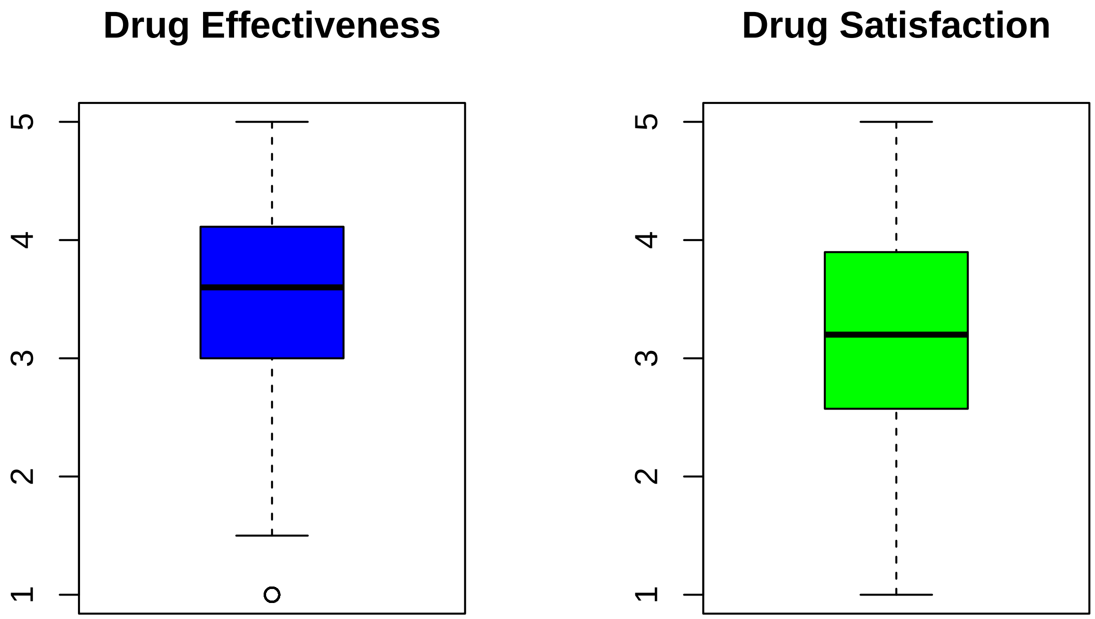
#Graphical analysis of the data Scatter Plot

```
scatter.smooth(x = data$Satisfaction, y = data$Effective, main = "Satisfaction vs Effectiveness")
points(x = data$Satisfaction, y = data$Effective, col = "green", pch = 16)
points(x = data$Satisfaction, y = data$Effective, col = rgb(0, 0, 1, alpha = 0.5), pch = 16)
```



#Graphical analysis of the data Box-Plot Of the data

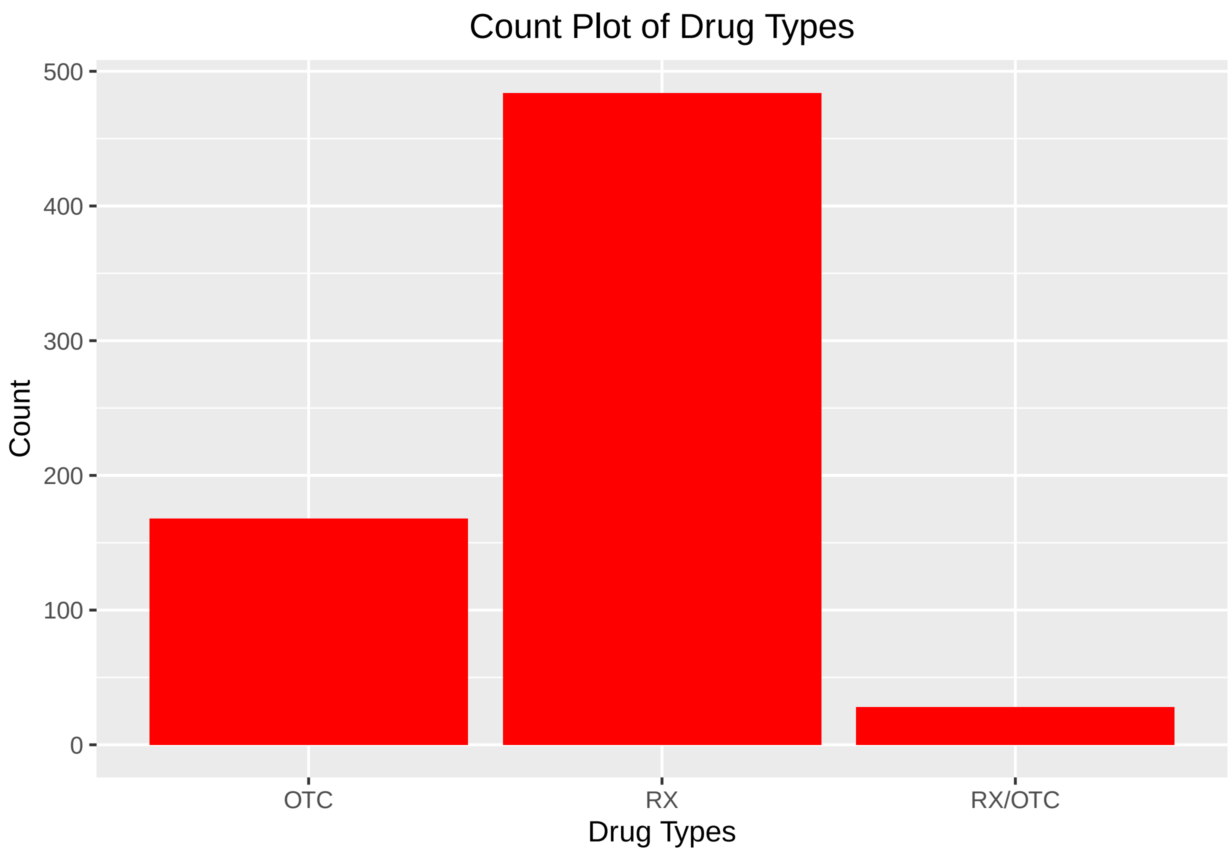
```
par(mfrow = c(1,2))
boxplot(data$Effective,main = "Drug Effectiveness", col = "blue")
boxplot(data$Satisfaction,main = "Drug Satisfaction", col = "green")
```



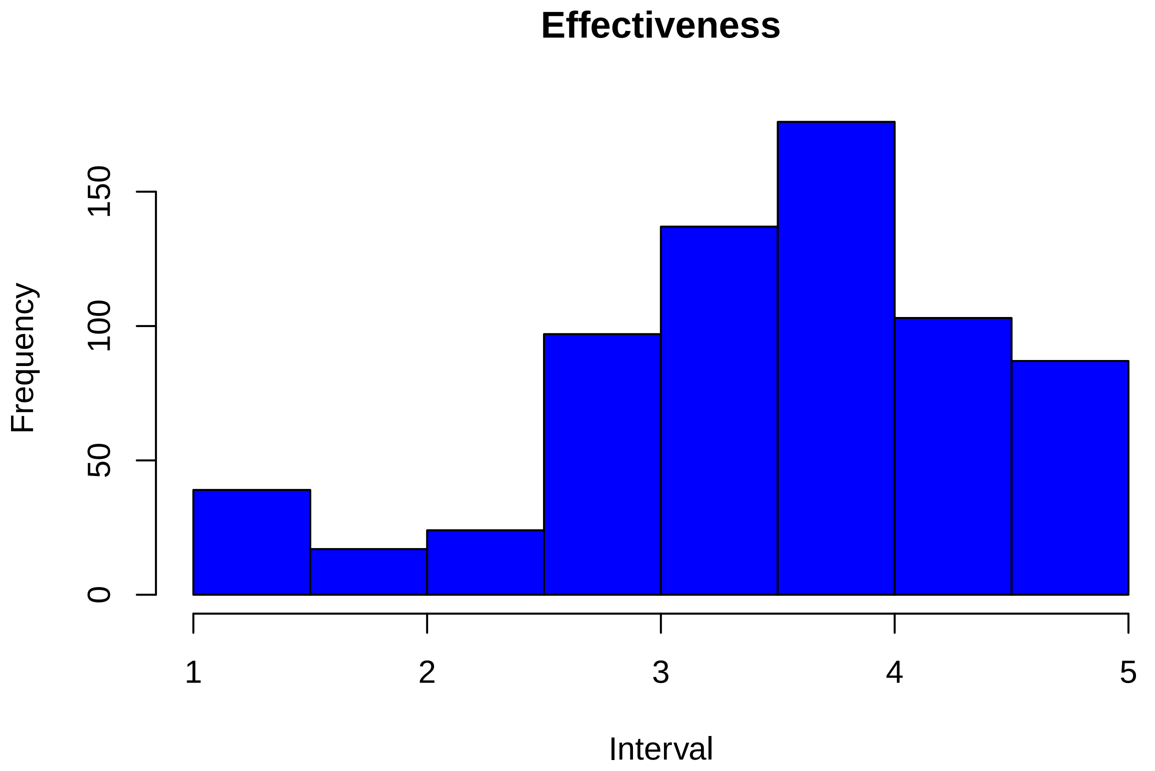
install.packages("moments") library(moments)

#Graphical analysis of the data Histogram

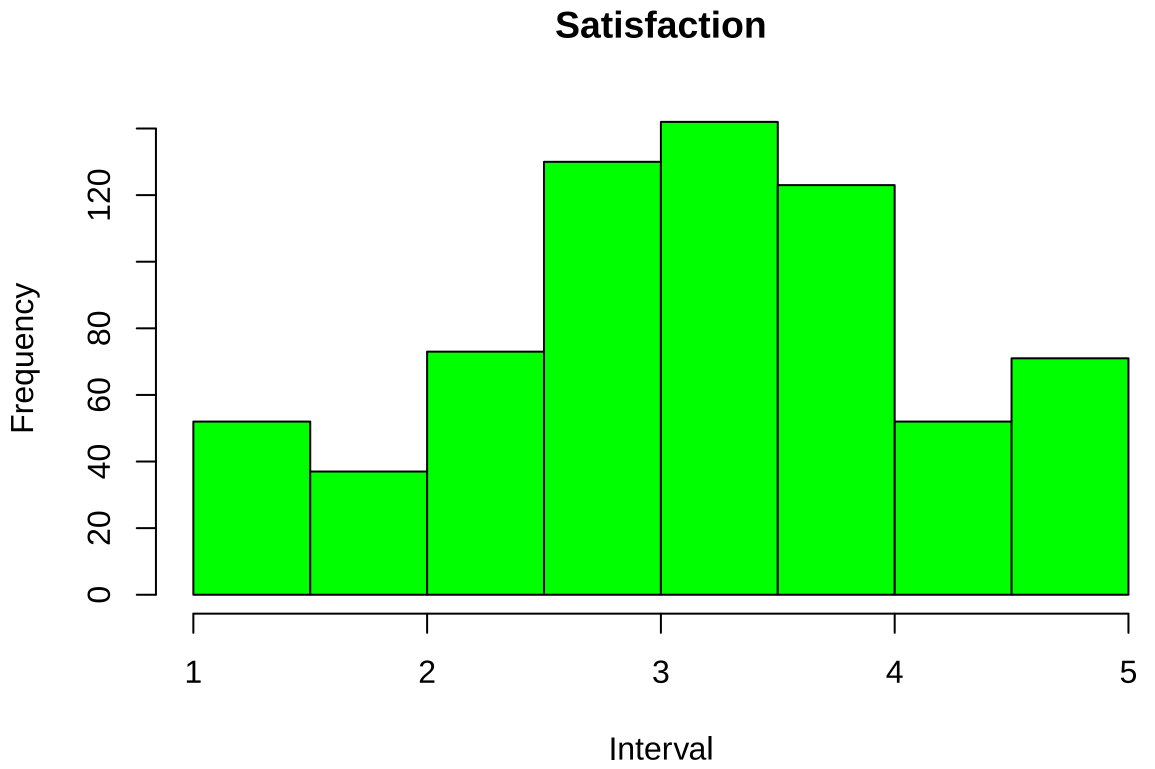
```
#Count for Drug Type
library(ggplot2)
plot <- ggplot(data, aes(x = Type)) +
  geom_bar(fill = "red") +
  labs(title = "Count Plot of Drug Types", x = "Drug Types", y = "Count") +
  theme(plot.title = element_text(hjust = 0.5))

# Display the plot
print(plot)
```



```
#Histogram for Effective
hist(data$Effective, col = "blue",
     xlab = "Interval", main = "Effectiveness")
```



```
#Histogram for Satisfaction
hist(data$Satisfaction, col = "green",
     xlab = "Interval", main = "Satisfaction")
```

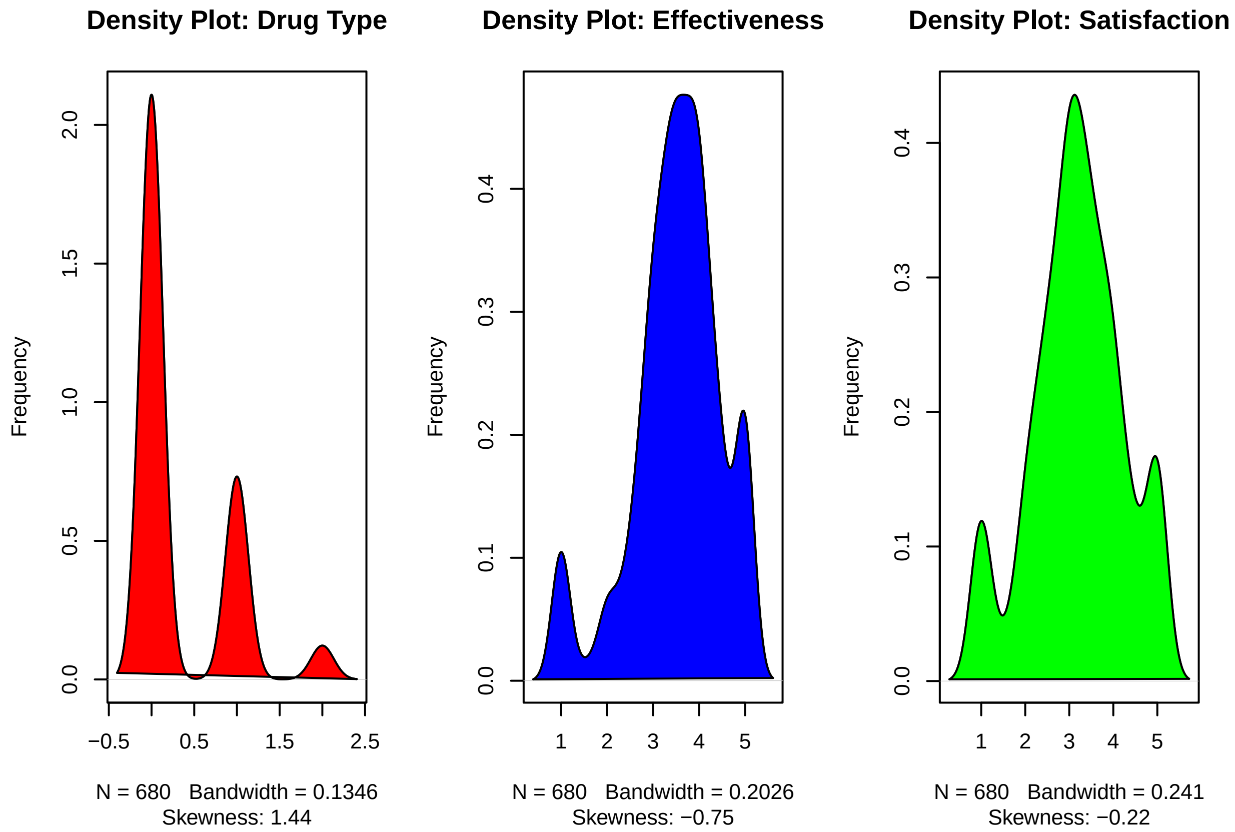


#Graphical analysis of the data Density Plot

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.2.3
```

```
par(mfrow=c(1, 3))
plot(density(data$Status), main="Density Plot: Drug Type", ylab="Frequency", sub=paste("Skewness:", rour
polygon(density(data$Status), col="red")

plot(density(data$Effective), main="Density Plot: Effectiveness", ylab="Frequency", sub=paste("Skewness
polygon(density(data$Effective), col="blue")

plot(density(data$Satisfaction), main="Density Plot: Satisfaction", ylab="Frequency", sub=paste("Skewnes
polygon(density(data$Satisfaction), col="green")
```



#Correlation Test between the variables

```
# Between Status and Effectiveness
cor(data$Status,data$Effective)
```

```
## [1] 0.05501989
```

```
# Between Status and Satisfaction
cor(data$Status,data$Satisfaction)
```

```
## [1] 0.1580571
```

#Shapiro-Wilk Test

```
# Perform Shapiro-Wilk test for Effective by Status
shapiro.test(data$Effective)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Effective
## W = 0.93879, p-value = 4.219e-16
```

```r
# Perform Shapiro-Wilk test for Satisfaction by Status
shapiro.test(data$Satisfaction)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Satisfaction
## W = 0.96768, p-value = 4.241e-11
```

#Kruskal-Wallis Test

```r
# Perform Kruskal-Wallis test for Effective by Status
kruskal.test(Effective ~ Status, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Effective by Status
## Kruskal-Wallis chi-squared = 17.096, df = 2, p-value = 0.0001939
```

```r
# Perform Kruskal-Wallis test for Satisfaction by Status
kruskal.test(Satisfaction ~ Status, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Satisfaction by Status
## Kruskal-Wallis chi-squared = 39.132, df = 2, p-value = 3.181e-09
```

#Calculating the Skewness and the Kurtosis for the Continous Variables

```r
# Calculate skewness
skew <- skewness(data$Satisfaction)
cat("The skewness of the Satisfaction variable is:", skew, "\n")
```

```
## The skewness of the Satisfaction variable is: -0.2169119
```

```r
# Calculate kurtosis
kurt <- kurtosis(data$Satisfaction)
cat("The kurtosis of the Satisfaction variable is:", kurt, "\n")
```

```
## The kurtosis of the Satisfaction variable is: -0.2403637
```

```r
# Calculate skewness
skew <- skewness(data$Effective)
cat("The skewness of the Effective variable is:", skew, "\n")
```

```
## The skewness of the Effective variable is: -0.7535869
```

```r
# Calculate kurtosis
kurt <- kurtosis(data$Effective)
cat("The kurtosis of the Effective variable is:", kurt, "\n")
```

```
## The kurtosis of the Effective variable is: 0.6951788
```

#Building the Generalized Linear Model Effectiveness and Satisfaction Combined

```r
model1 <- glm(Status ~ Effective + Satisfaction, data = data)
summary(model1)
```

```
##
## Call:
## glm(formula = Status ~ Effective + Satisfaction, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6669  -0.3514  -0.2676   0.5223   1.7378
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.24173    0.07897   3.061  0.00229 **
## Effective    -0.18926    0.04321  -4.380 1.38e-05 ***
## Satisfaction  0.23644    0.04005   5.904 5.61e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2888043)
##
##     Null deviance: 206.21  on 679  degrees of freedom
## Residual deviance: 195.52  on 677  degrees of freedom
## AIC: 1090.2
##
## Number of Fisher Scoring iterations: 2
```

#Proceeding ahead with Mutlinomial Logistic Regression owing to the high AIC by GLM

#Building the Multinomial Logistic Regression Model Effectiveness and Satisfaction Combined

```r
library(nnet)
model2 <- multinom(Status ~ Effective + Satisfaction, data = data)
```

```
## # weights:  12 (6 variable)
## initial  value 747.056356
## iter  10 value 463.068247
## iter  20 value 460.538397
## iter  20 value 460.538394
## iter  20 value 460.538394
## final  value 460.538394
## converged
```

```
summary(model2)
```

```
## Call:
## multinom(formula = Status ~ Effective + Satisfaction, data = data)
##
## Coefficients:
##   (Intercept) Effective Satisfaction
## 1   -1.918732 -1.018603     1.359812
## 2   -2.176566 -1.125653     1.014652
##
## Std. Errors:
##   (Intercept) Effective Satisfaction
## 1   0.3810441 0.2491760    0.2321839
## 2   0.6419704 0.4576465    0.4438286
##
## Residual Deviance: 921.0768
## AIC: 933.0768
```

#Multinomial Logistic Regression performs better when compared to GLM on Residual Deviance and AIC

#Building the Multinomial Logistic Regression Model Between Satisfaction and Status

```
library(nnet)
model3 <- multinom(Status ~ Satisfaction, data = data)
```

```
## # weights:  9 (4 variable)
## initial  value 747.056356
## iter  10 value 471.863160
## final  value 471.862290
## converged
```

```
summary(model3)
```

```
## Call:
## multinom(formula = Status ~ Satisfaction, data = data)
##
## Coefficients:
##   (Intercept) Satisfaction
## 1   -2.831545   0.53271944
## 2   -2.978803   0.04181026
##
## Std. Errors:
##   (Intercept) Satisfaction
## 1   0.3440076   0.09603717
## 2   0.6241925   0.19116386
##
## Residual Deviance: 943.7246
## AIC: 951.7246
```

##Building the Multinomial Logistic Regression Model Between Effectiveness and Status

```r
library(nnet)
model4 <- multinom(Status ~ Effective, data = data)
```

```
## # weights:  9 (4 variable)
## initial  value 747.056356
## iter  10 value 483.491500
## final  value 483.491484
## converged
```

```r
summary(model4)
```

```
## Call:
## multinom(formula = Status ~ Effective, data = data)
##
## Coefficients:
##   (Intercept)  Effective
## 1   -2.118127  0.2944319
## 2   -2.245028 -0.1788270
##
## Std. Errors:
##   (Intercept) Effective
## 1   0.3824528 0.1014545
## 2   0.6561624 0.1896376
##
## Residual Deviance: 966.983
## AIC: 974.983
```