BT1101

Lab 3: Statistical Measures, Probability Distributions and Data Modelling

Installing and loading packages

```
# install required packages if you have not (suggested packages: dplyr, tidyr, rpivotTable, knitr, psych)
# install.packages("dplyr") #only need to run this code once to install the package
# load required packages
# library("xxxx")
library("dplyr") #need to call the library before you use the package
```

```
library("tidyr")
library("rpivotTable")
library("knitr")
library("psych")
```

Default Requirements for Tables/Charts

- Tables must have appropriate titles and column names
- Charts must have appropriate titles, axis labels, and legend (where necessary)
- Pie charts should have appropriate titles and slices should be labeled with the category name and percentage value.

These default requirements should be followed unless specified otherwise in the question.

Part 1

Tutorial 4 Part 1 (For lab session)

- Dataset required: `Sales Transactions.xlsx`

`Sales Transactions.xlsx` contains the records of all sale transactions for a day, July 14. Each of the column is defined as follows:

- `CustID` : Unique identifier for a customer
- `Region` : Region of customer's home address
- `Payment` : Mode of payment used for the sales transaction
- `Transction Code` : Numerical code for the sales transaction
- `Source` : Source of the sales (whether it is through the Web or email)
- `Amount` : Sales amount
- `Product` : Product bought by customer
- `Time Of Day` : Time in which the sale transaction took place.

As the business analytics analyst of the company, you have been tasked to help the store manager develop dashboard that will enable him to gain better insights of the data.

Loading datasets into R

```
#put in your working directory folder pathname ()

#import excel file into RStudio
library(readxl)
setwd("C:/nbox/Soc Acad Courses/AY2022 BT1101/Data")
#import xlsx file into RStudio
ST <- read_excel("Sales Transactions.xlsx", col_types = c("numeric", "text", "text", "numeric", "text", "numeric", "text",
"date"), skip = 2)
head(ST)
```

```
## # A tibble: 6 × 8
##   `Cust ID` Region Payment `Transaction Code` Source Amount Product
##       <dbl> <chr>  <chr>                <dbl> <chr>   <dbl> <chr>
## 1     10001 East   Paypal            93816545 Web      20.2 DVD
## 2     10002 West   Credit            74083490 Web      17.8 DVD
## 3     10003 North  Credit            64942368 Web      24.0 DVD
## 4     10004 West   Paypal            70560957 Email    23.5 Book
## 5     10005 South  Credit            35208817 Web      15.3 Book
## 6     10006 West   Paypal            20978903 Email    17.3 DVD
## # … with 1 more variable: `Time Of Day` <dttm>
```

Coding Practice

Q1.(a) Customer Dashboard

The manager would like to have a better understanding of the customer profiles. He would like the customer dashboard to be able to display in charts and tables, the following:

- i. frequency distribution for the regions the customers are from
- ii. frequency distribution for the payment mode used by the customers

He would like you to use shades of blue for the charts. He would also like to have your interpretation of the tables and charts generated. Write your observation in the space below.

Q1(a)i- Customer Dashboard

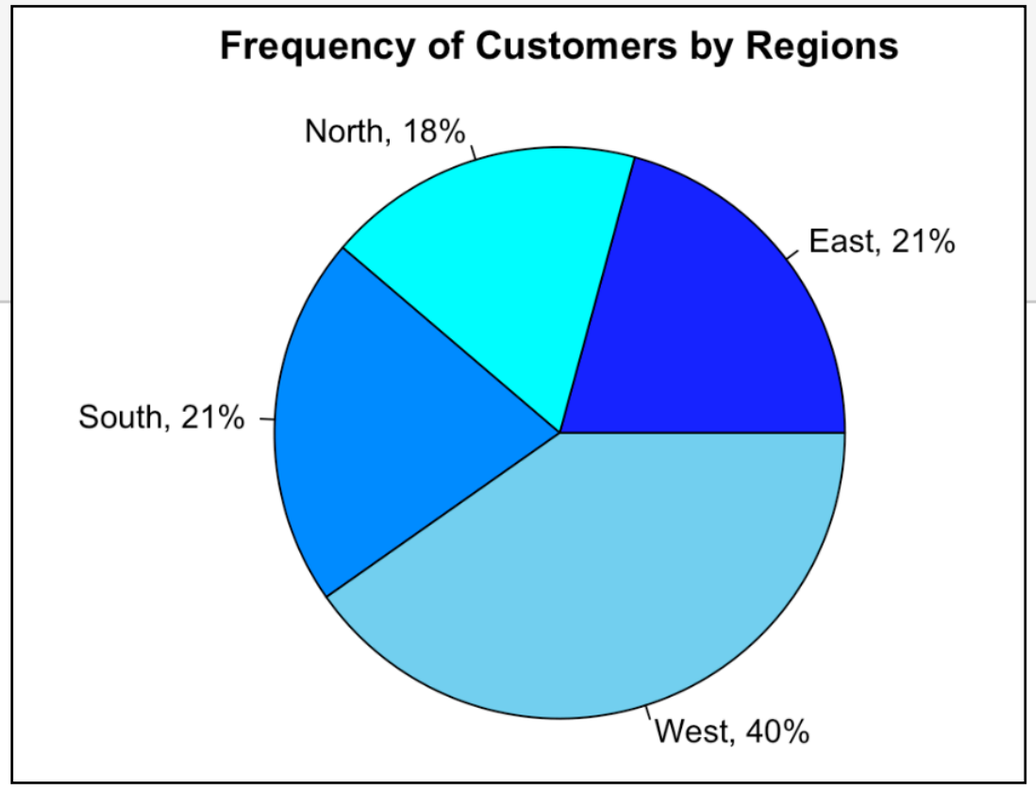
Frequency distribution for regions customers are from: Single categorical variable, so we use a pie chart (or bar chart)

```
#Pie chart for Region (Barchart is also appropriate)
Freq.reg<-ST %>% count(`Region`)
kable(Freq.reg, caption = "Frequency of Customers by Region")
```

```
slice.reg <- Freq.reg$n
reg.piepercent <- 100*round(Freq.reg$n/sum(Freq.reg$n),2)
label<-Freq.reg$Region
label<-paste(label,",",sep="")
label<-paste(label,reg.piepercent) #default of sep=" "
label<-paste(label,"%",sep="")
pie(slice.reg,
    labels=label,
    col=c("blue","cyan","dodgerblue", "skyblue"),
    radius=1,
    main="Frequency of Customers by Regions")
```

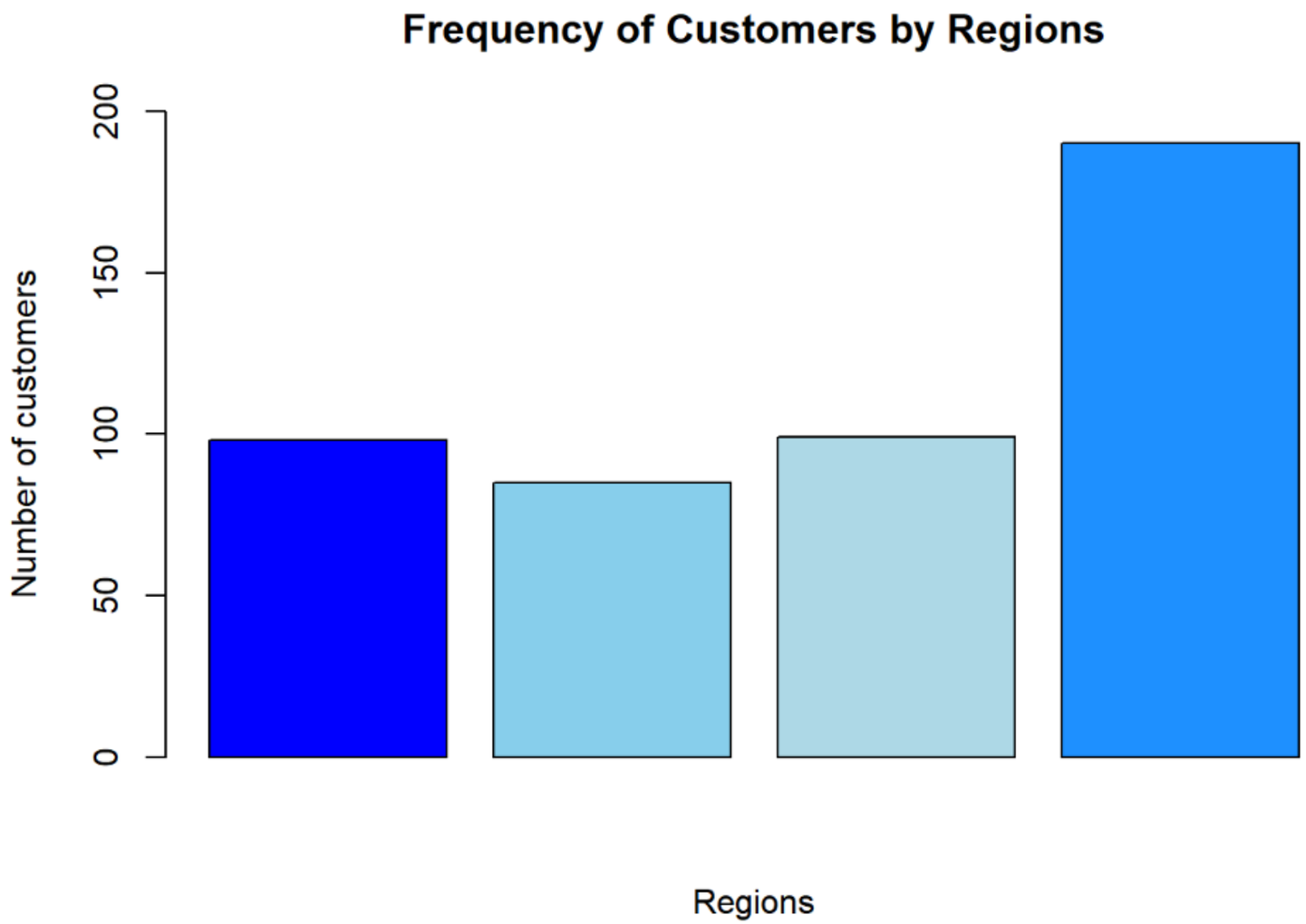
Frequency of Customers by Region

Region	n
East	98
North	85
South	99
West	190



Q1(a) - Customer Dashboard

```
barplot(Freq.reg$n, ylab="Number of customers", xlab="Regions", col=c("blue","skyblue","lightblue","dodgerblue"),ylim= c(0,200), main = "Frequency of Customers by Regions")
```



What is your interpretation of the tables and charts generated?

Q1(a)ii - Customer Dashboard

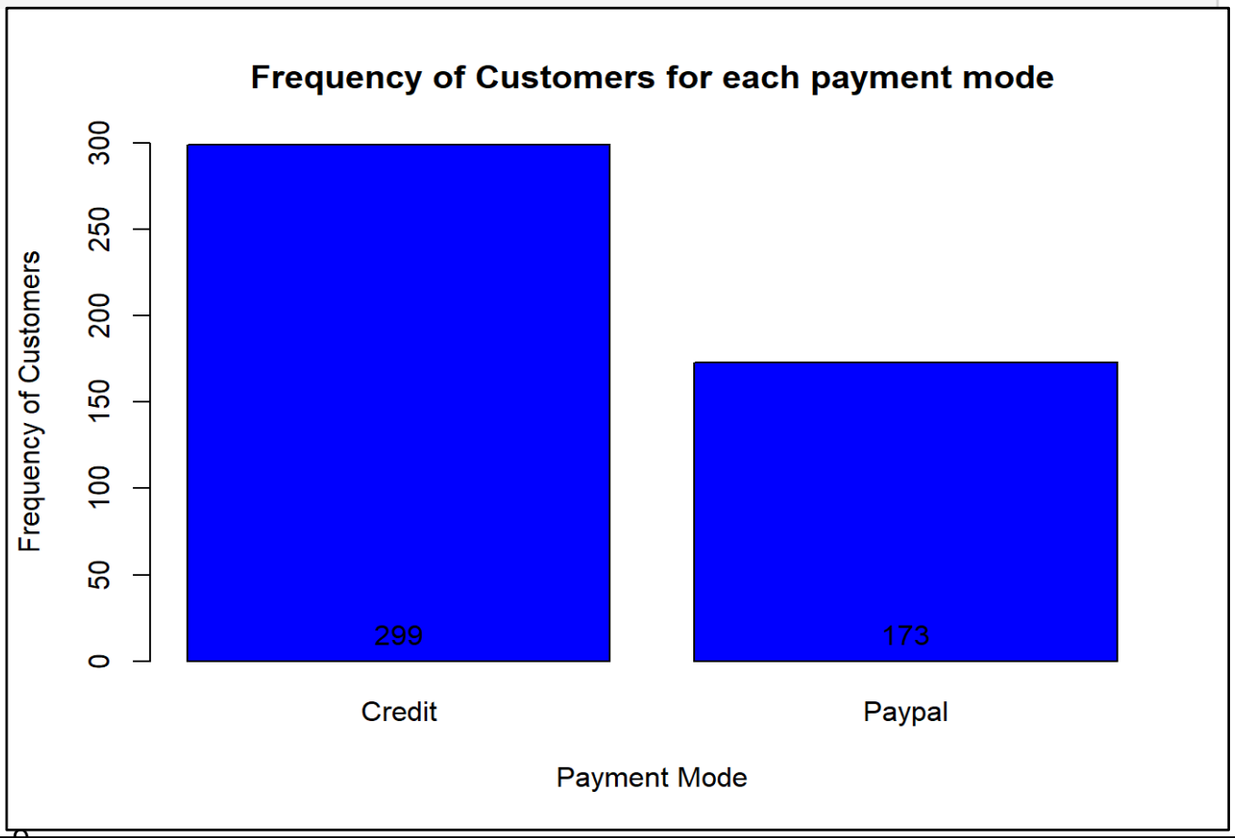
Frequency distribution for payment mode used by the customers: Single categorical variable, so we use a pie chart (or bar chart)

```
# Barchart for Payment (Pie chart is also appropriate. Here we provide an eg of each.)
Freq.pay<-ST %>% count(`Payment`)
kable(Freq.pay, caption = "Frequency of Customers for each payment mode")
```

```
bp<-barplot(Freq.pay$n, ylab="Frequency of Customers", ylim=c(0,300), names.arg= Freq.pay$Payment, xlab="Payment Mode", main
="Frequency of Customers for each payment mode",col = "blue")
# If label is required for the bars
text(bp,0, Freq.pay$n, pos=3)
```

Frequency of Customers for each payment mode

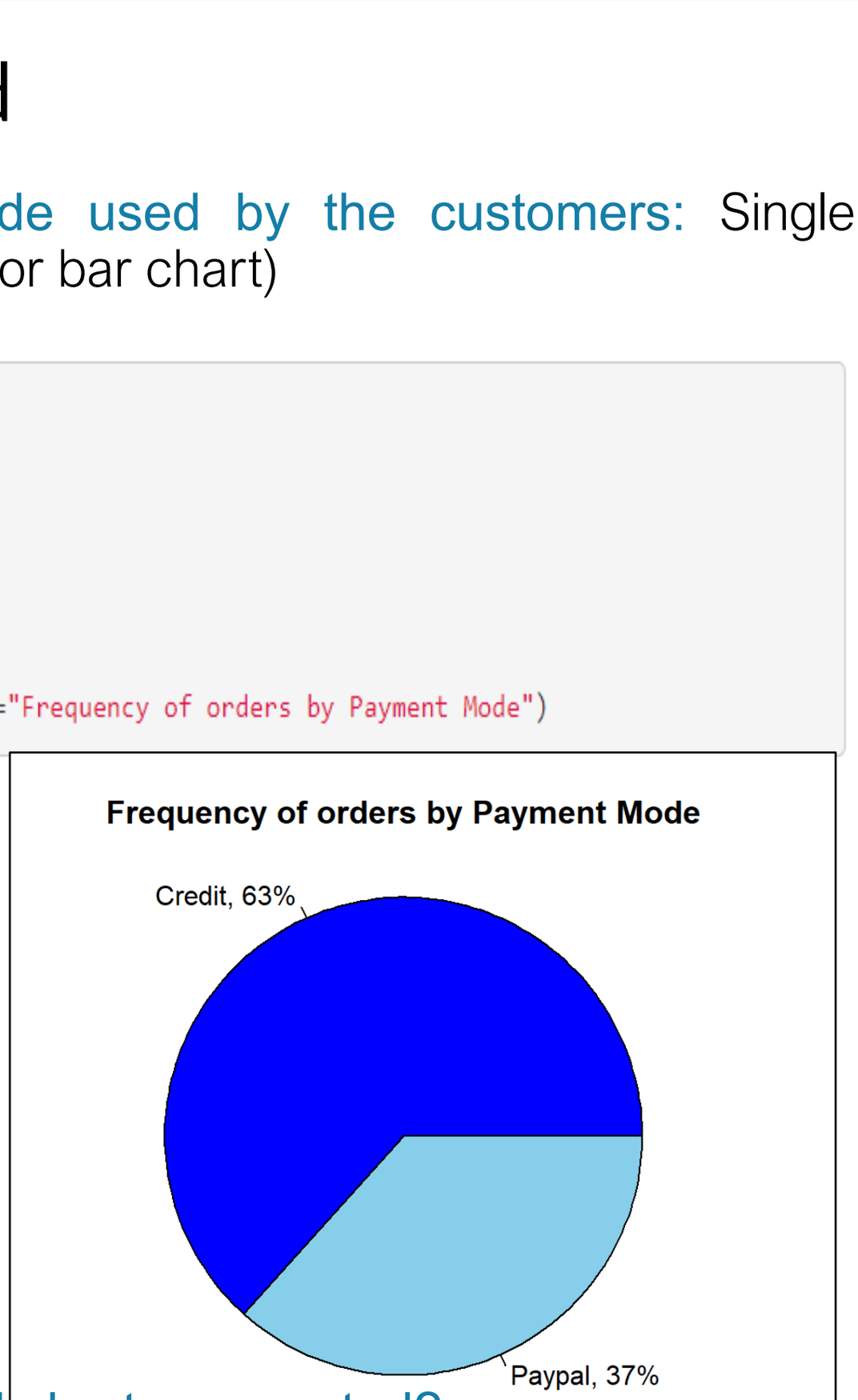
Payment	n
Credit	299
Paypal	173



Q1(a)ii - Customer Dashboard

Frequency distribution for payment mode used by the customers: Single categorical variable, so we use a pie chart (or bar chart)

```
slice.pay <- Freq.pay$n
pay.piepercent <- 100*round(Freq.pay$n/sum(Freq.pay$n),2)
label<-Freq.pay$Payment
label<-paste(label,",",sep="")
label<-paste(label,pay.piepercent) #default of sep=" "
label<-paste(label,"%",sep="")
pie(slice.pay,labels=label, col=c("blue","skyblue"),radius=1, main="Frequency of orders by Payment Mode")
```



What is your interpretation of the tables and charts generated?

Coding Practice

Q1.(b) Sales Transaction Analyses Dashboard

The manager would also like to have a dashboard to be able to visualize the sales `Amount` data better.

- i. First, generate the descriptive statistics for `Amount` in a table. The manager would like to include only these statistics: n (or number of observations), mean, sd, median, skew, kurtosis. (Discuss what these statistics tell you about the distribution of `Amount`. Is it normally distributed?)
- ii. Plot the histogram, density plot and normal Q-Q plot for `Amount`. Then conduct the appropriate goodness of fit test to confirm if the variable is normally distributed. [Note: Typically you can choose which plot to plot that will enable you to make a better judgement]
- iii. The manager is concerned about potential outliers in the data. Can you help to identify if any outliers for `Amount` exists?
- iv. The manager suspects that the sales `Amount` may differ for transactions involving `Book` versus `DVD`. Could you generate the table and chart for him to be able to compare the mean and standard deviations of `Amount` for books versus dvds? Describe what you can observe from the chart.
- v. Perform the outlier analyses separately for books and dvds. What observations can you make now? Would you remove any of the outliers?

Q1(b) - Sales Transaction Analyses Dashboard

i. First, generate the **descriptive statistics for** Amount in a table: n, mean, sd, median, skew, kurtosis. Discuss what these statistics tell you about the distribution of Amount.

```
# Generate Descriptive stats for Amount
tab.1b<-describe(ST$Amount)
tab.1b$range <- tab.1b$trimmed <- tab.1b$mad <- tab.1b$se <- tab.1b$min<-tab.1b$max <-NULL # remove columns not needed
tab.1b$vars[1]<-"Amount"
kable(tab.1b, row.names = FALSE, caption = "Descriptive Statistics for `Amount`")
```

Descriptive Statistics for `Amount`

vars	n	mean	sd	median	skew	kurtosis
Amount	472	39.94581	57.32009	20.605	2.596053	5.080512

Alternative:

```
# explicitly name package that function is from
tab.1b<-psych::describe(ST$Amount)
```

Q1(b) - Sales Transaction Analyses Dashboard

i. First, generate the **descriptive statistics for** Amount in a table: mean, sd, median, skew, kurtosis.

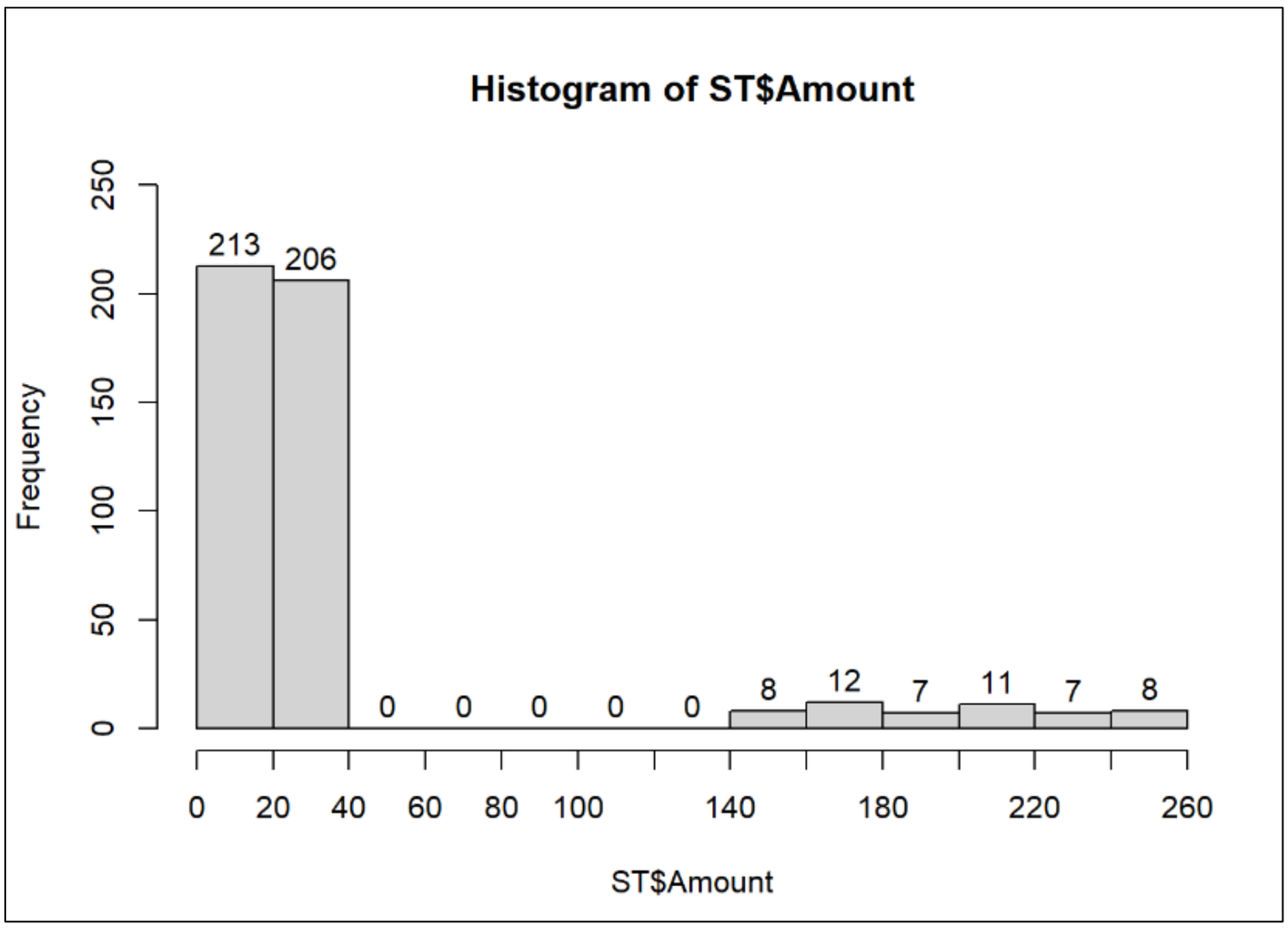
```
# with dplyr
ST %>%
  summarise(
    vars="Amount",
    n=n(),
    mean=mean(Amount),
    sd=sd(Amount),
    median=median(Amount),
    skew=skew(Amount),
    kurtosis=kurtosi(Amount)) %>%
mutate(across(where(is.double), round, 2)) %>% # specify no. decimal places
kable(row.names=FALSE, caption="Descriptive Statistics for `Amount`")
```

Instead of using **describe** from the psych package, we can also calculate the required statistics using **summarise** from dplyr.

Q1(b) - Sales Transaction Analyses Dashboard

ii. Plot the histogram for **Amount** and conduct the appropriate **goodness of fit test to confirm if it is normally distributed**.

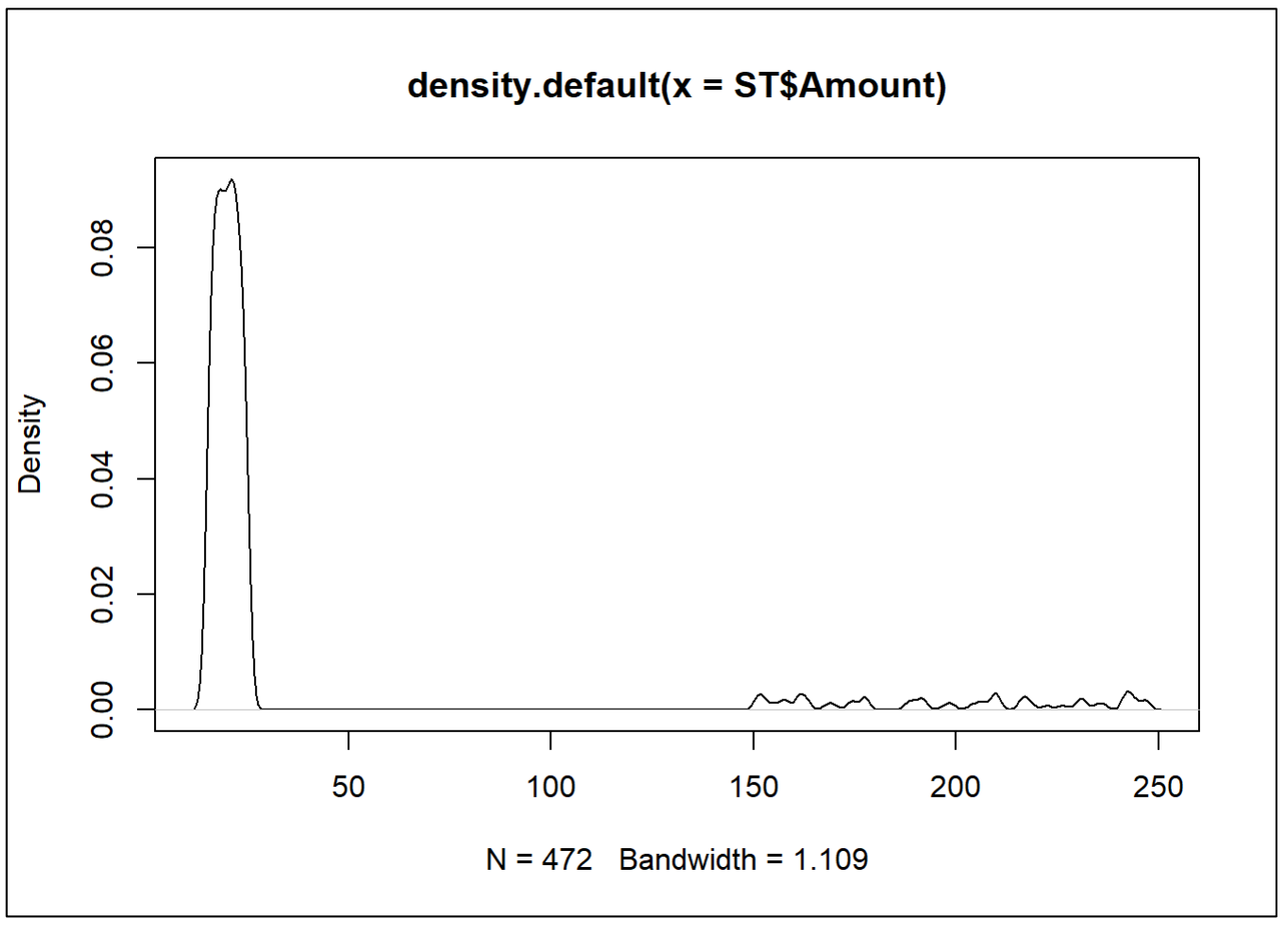
```
# (ii)
# Histogram
H<-hist(ST$Amount, ylim=c(0,250), labels = TRUE, xaxp=c(0,260, 13))
```



Q1(b) - Sales Transaction Analyses Dashboard

ii. Plot the density plot for **Amount** and conduct the appropriate **goodness of fit test to confirm if it is normally distributed**.

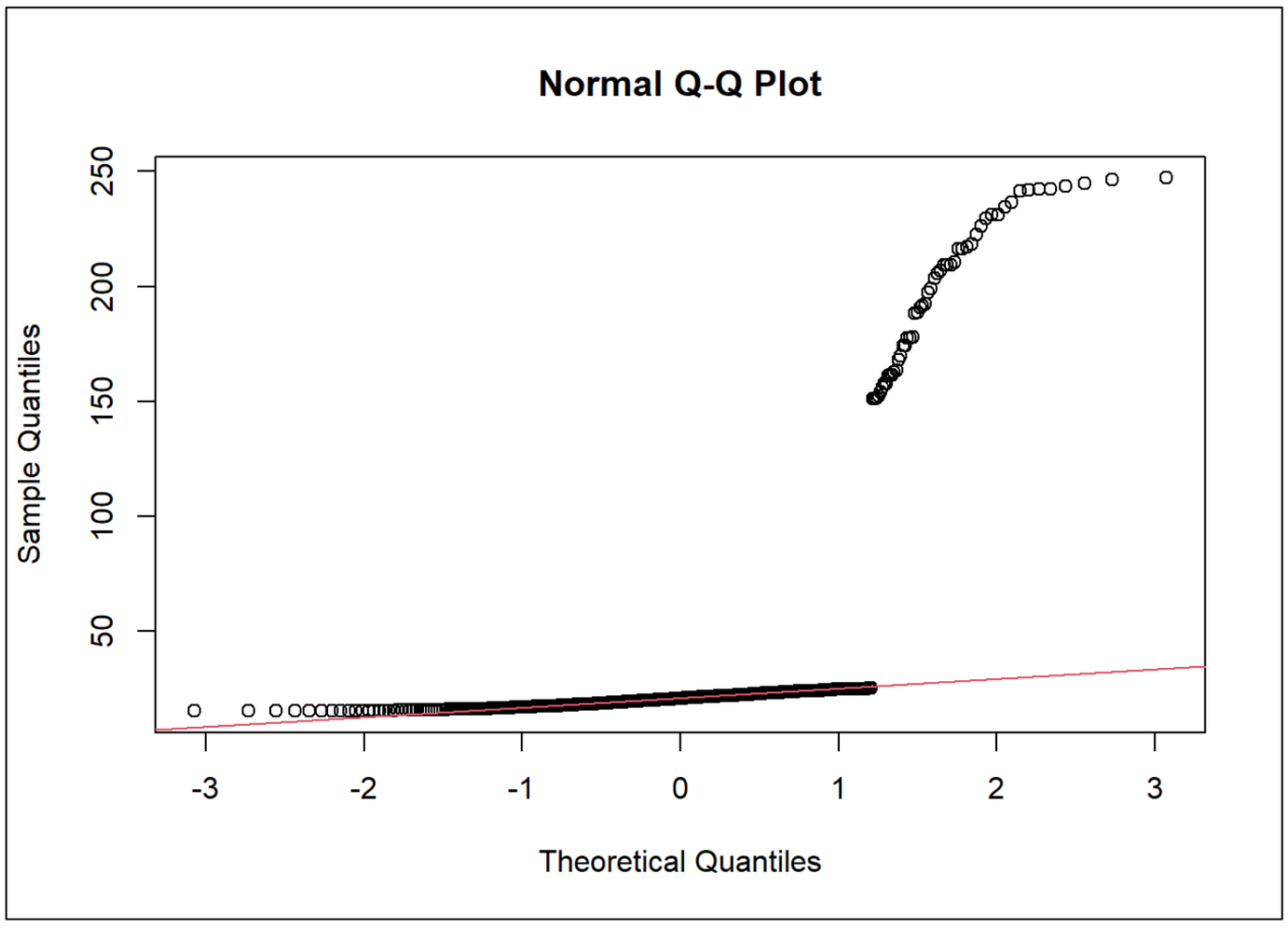
```
# density plot for Amount
plot(density(ST$Amount))
```



Q1(b) - Sales Transaction Analyses Dashboard

ii. Plot the normal Q-Q plot for **Amount** and conduct the appropriate **goodness of fit test to confirm if it is normally distributed**.

```
# normal Q-Q plot for Amount
qqnorm(ST$Amount)
qqline(ST$Amount, col=2)
```



Q1(b) - Sales Transaction Analyses Dashboard

ii. Conduct the appropriate **goodness of fit test to confirm if it is normally distributed**.

```
# Shapiro-Wilk Test
shapiro.test(ST$Amount)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ST$Amount
## W = 0.42617, p-value < 2.2e-16
```

- **W** is the test statistic for the Shapiro-Wilk test
- Null hypothesis H_0: Data was drawn from a normally distributed population
- We typically set **cutoff/thresholds** for the p-value to determine statistical significance.
  - 0.05 is a common value.
- Interpretation: Since $p<0.05$, the probability of the available data is less than 5% given that the null hypothesis is true.
  - There is evidence that the data is **not normally distributed**.

Q1(b) - Sales Transaction Analyses Dashboard

iii. The manager is concerned about potential outliers in the data. Can you help to **identify if any outliers for** Amount **exist**?

We can use visual aids e.g. box plots to help identify possible outliers.

Outlier analyses can be done in a few ways:

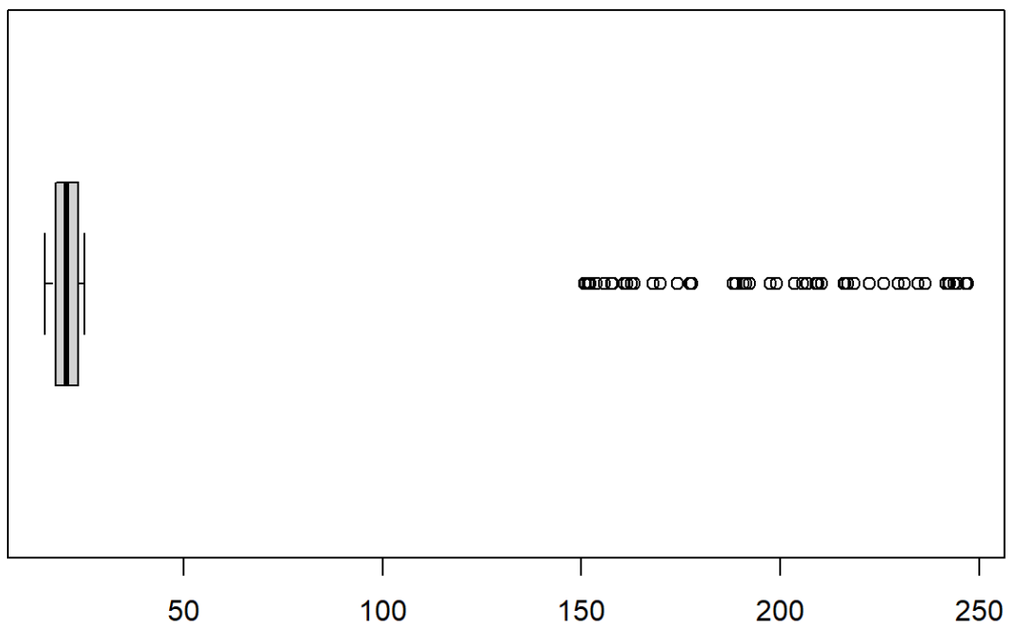
1. Visual inspection
2. Boxplots (assumes **normally distributed** data)
3. Rules of thumb (assumes **normally distributed** data)

When data is **skewed**, visual inspection should be used with charts such as histograms for outlier identification.

Q1(b) - Sales Transaction Analyses Dashboard

iii. The manager is concerned about potential outliers in the data. Can you help to identify if any outliers for Amount exist?

```
# Boxplot can be plotted to show students for a comparison with the histogram.
boxplot(ST$Amount, range=3, horizontal = TRUE)
```



- The **range** argument determines how far the plot whiskers extend out of the box.
- What do you observe? Do you think the data contains outliers?

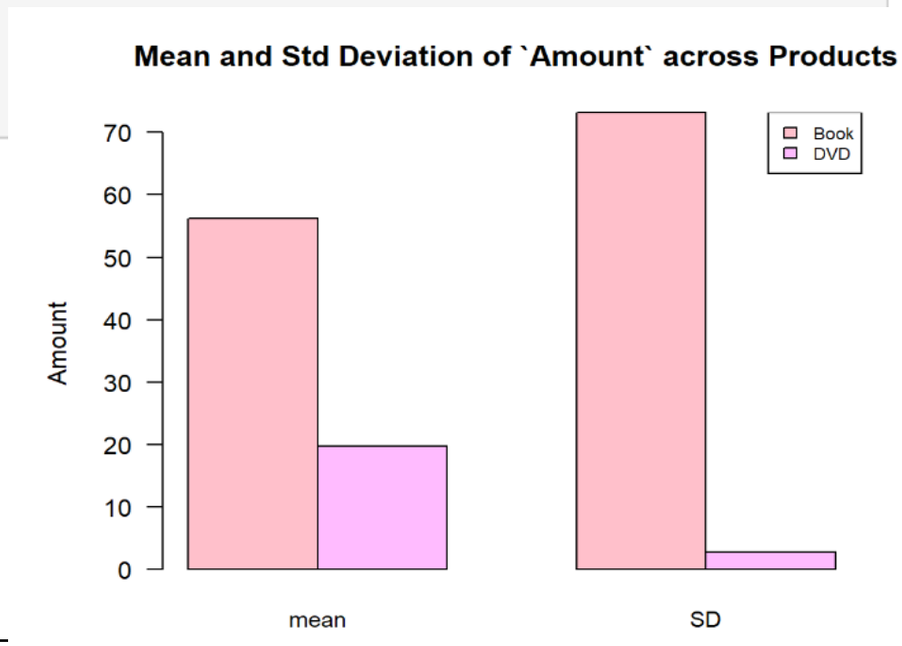
Q1(b) - Sales Transaction Analyses Dashboard

iv. The manager suspects that the sales Amount may differ for transactions involving Book versus DVD. Could you generate the table and chart for him to be able to compare the mean and standard deviations of Amount for books versus dvds? Describe what you can observe from the chart.

```
#
tab.1b2<- ST %>% group_by(`Product`) %>% summarise(mean=mean(Amount), SD=sd(Amount))
kable(tab.1b2)
```

Product	mean	SD
Book	56.21559	73.15149
DVD	19.82062	2.81961

```
#plot grouped barplot
par(mar=c(5,10,4,2)) # default plot margin is (5,4,4,2), I'm adding a bigger left margin for the barchart
bar.1b2<-as.matrix(tab.1b2[,c(2:3)])
col.1b2<-c("pink","plum1")
barplot(bar.1b2,  beside= TRUE, col =col.1b2, main=" Mean and Std Deviation of `Amount` across Products", cex.names=0.9,  la
s=1,  ylab="Amount")
legend("topright", cex=0.7, fill=col.1b2, tab.1b2$Product)
```



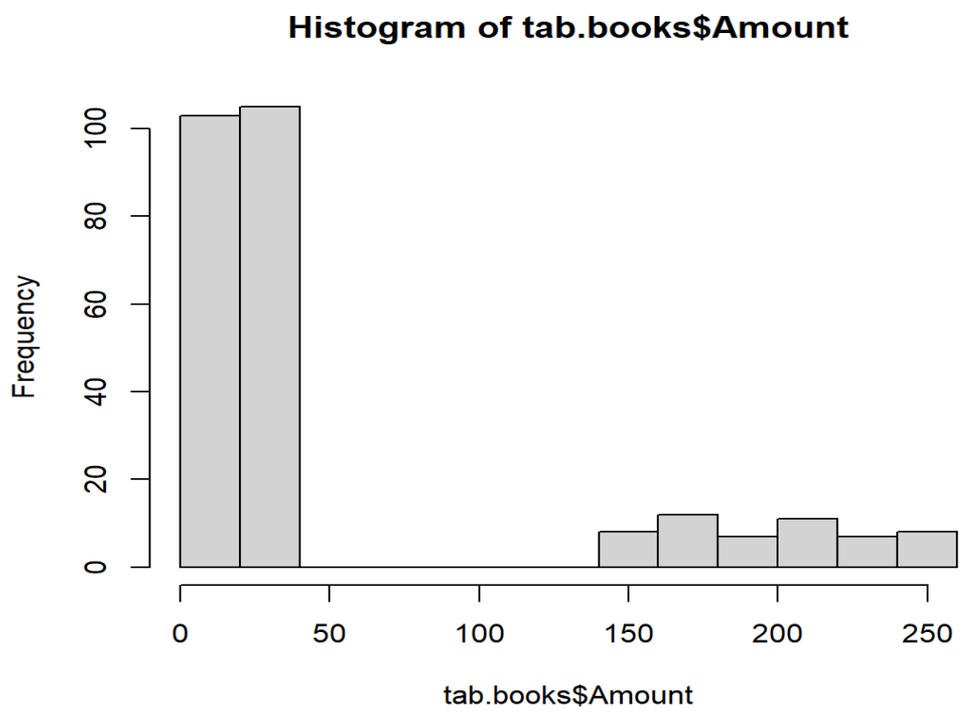
Q1(b) - Sales Transaction Analyses Dashboard

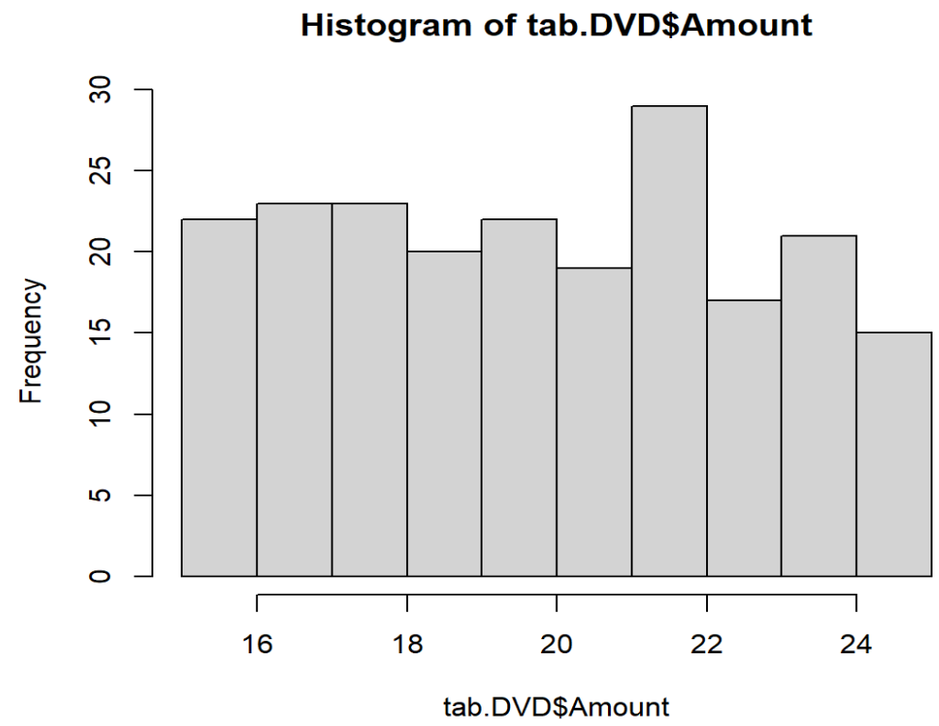
v. Perform the **outlier analyses** separately for books and dvds. What observations can you make now? Would you remove any of the outliers?

```
# first we split the data
tab.books<-ST%>%filter(Product=="Book")
tab.DVD<-ST%>%filter(Product=="DVD")

# Then we plot the histogram for each data to see if data is normally distributed
hist(tab.books$Amount)
```

```
hist(tab.DVD$Amount)
```





Q1(b) - Sales Transaction Analyses Dashboard

- Since both data sets are **not normally distributed**, we can just use visual inspection for outlier detection.

- We can see that there are still two groups in the Books data but there isn't for DVD data. So we can conclude that there are **no outliers for DVD data**.

- In the case of books, there are quite a number of sales with higher sales amount. Therefore they are **unlikely to be outliers**.

  - Discuss with the book store manager reveals that higher sales amount is due to the sales of rare/collector item books that tend to cost more. Note: They are not outliers.

  - To deal with "outlier" here, one way is to analyse normal books and rare/collector books separately. This is something that needs to be discussed with the mgr.

manager!

Q1(b) - Sales Transaction Analyses Dashboard

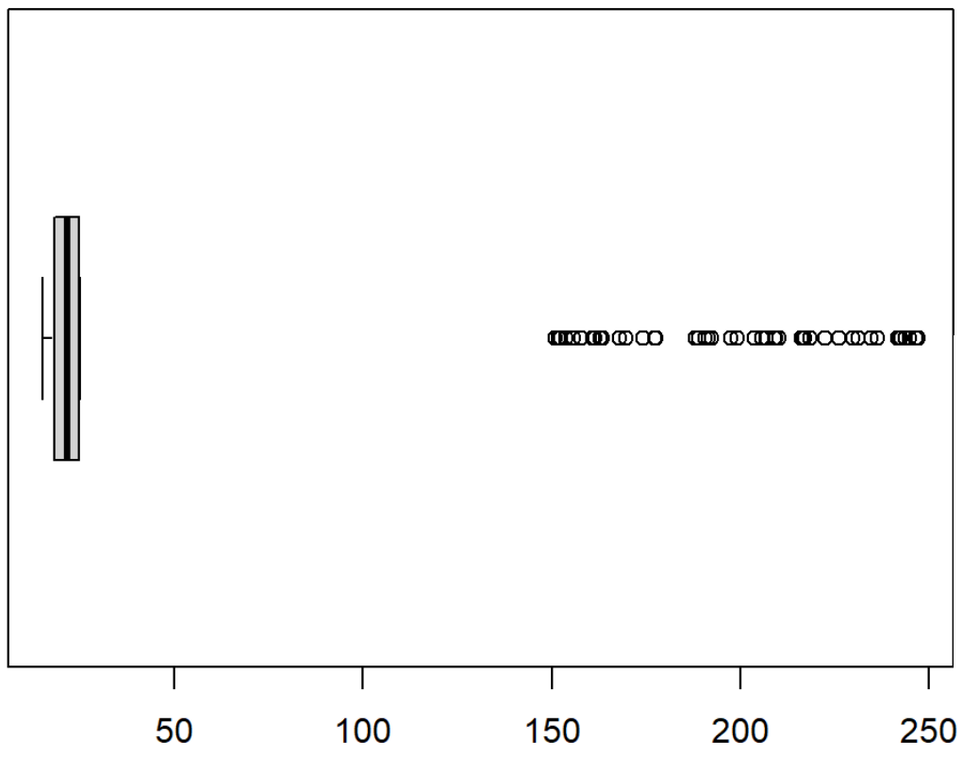
Note: If the data is normally distributed and we are using boxplot to extract the outliers,

We can apply the following technique:

```
# If the data is normally distributed and we are using boxplot to extract the outliers, we can apply the following techniqu
e:
boxplot.bk<-boxplot(tab.books$Amount, horizontal=TRUE, range=3)
```

```
boxplot.bk$out
```

the out variable from the output of the boxplot function will enable us to extract the points that lie beyond the extremes of the whiskers




```
##  [1] 177.72 151.67 205.58 206.80 217.00 150.99 209.51 229.73 157.76 216.37
## [11] 174.25 209.37 174.18 236.49 155.91 234.63 190.81 177.32 241.77 192.41
## [21] 242.52 226.15 216.20 161.46 243.70 210.38 161.50 209.20 191.43 241.65
## [31] 242.40 157.86 222.38 188.85 231.23 244.75 162.74 188.16 246.67 177.30
## [41] 203.72 150.86 199.18 197.43 153.83 160.78 169.79 152.27 218.60 163.37
## [51] 231.23 247.14 168.10
```

Coding Practice

Q1.(c) Checking Correlation

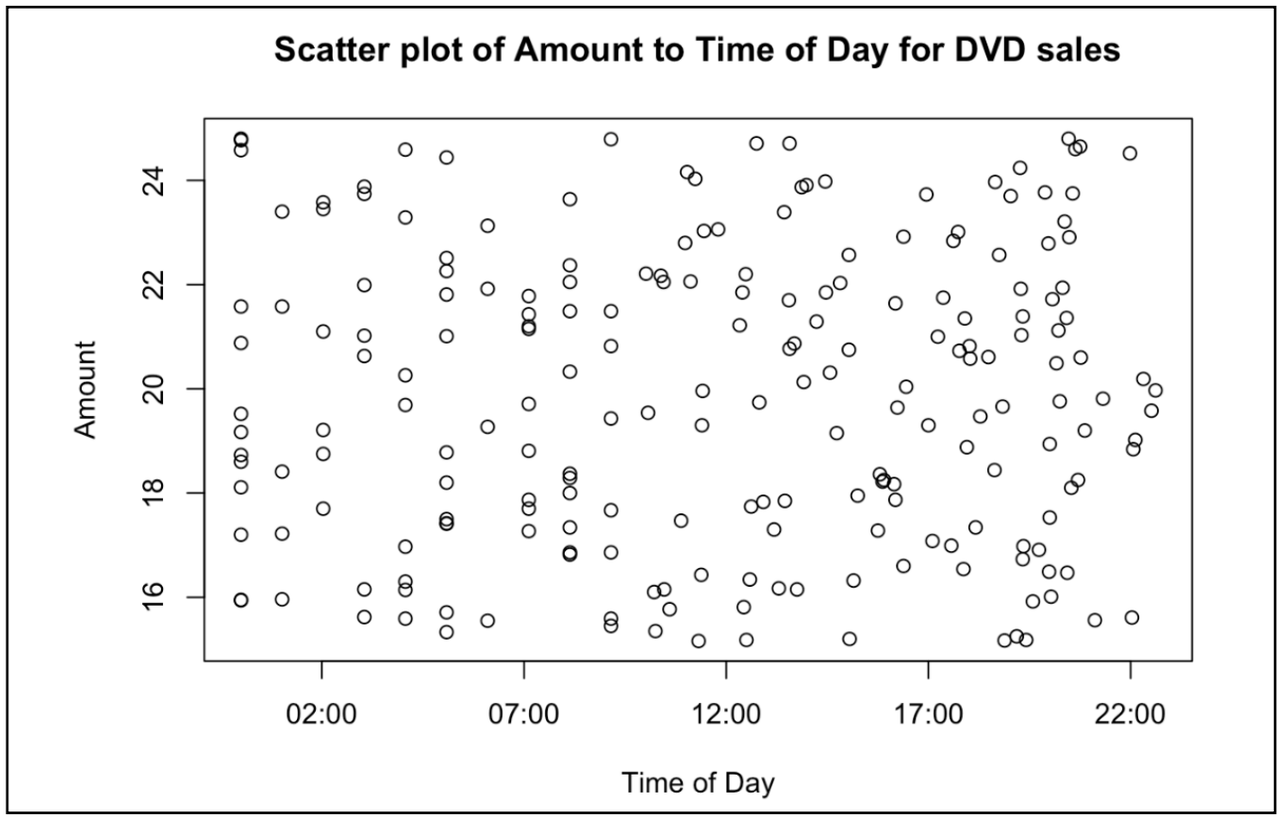
The manager would like to check if the sales `Amount` for DVD has any correlation with `Time of the Day`.

- i. Plot the appropriate chart and provide the statistical measure to help the manager assess this.
- ii. Type your interpretation for the manager in the space below.

Q1(c) - Checking Correlation

i. Plot the **appropriate chart** and provide the **statistical measure** to help the manager assess if the sales amount for DVD has any correlation with time of the day.

```
##(i)
plot(x=tab.DVD$`Time Of Day`,
     y=tab.DVD$Amount,
     main="Scatter plot of Amount to Time of Day for DVD sales",
     xlab="Time of Day",
     ylab = "Amount")
```



Q1(c) - Checking Correlation

i. Plot the **appropriate chart** and provide the **statistical measure** to help the manager assess if the sales amount for DVD has any correlation with time of the day.

```
cor(as.numeric(tab.DVD$`Time Of Day`), tab.DVD$Amount) # need to highlight that Time of Day is not numeric data so it needs to be converted first before using the cor function
```

```
## [1] 0.03188728
```

```
corr.test(as.numeric(tab.DVD$`Time Of Day`), tab.DVD$Amount)
```

```
## Call:corr.test(x = as.numeric(tab.DVD$`Time Of Day`), y = tab.DVD$Amount)
## Correlation matrix
## [1] 0.03
## Sample Size
## [1] 211
## These are the unadjusted probability values.
##   The probability values  adjusted for multiple tests are in the p.adj object.
## [1] 0.65
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

Q1.(d) Computing proportions and probability

The manager would like to use the existing data to compute the following:

- i. Proportion of `Book` sales transactions that have `Amount` greater than $60.
- ii. Proportion of `DVD` sales transactions that are from the Web.

Assume that we do not have this dataset that you are working with. Instead we are told the DVD sales Amount is normally distributed with a mean of $20 and standard deviation of $4. What is the probability of DVD sales amount being great than $25?

Q1(d) - Computing proportions and probability

i. Proportion of **Book** sales transactions that have **Amount** greater than $60

```
# i
df.book <- ST %>% filter(Product =="Book")
df.book60 <- df.book %>% filter(Amount>60)
nrow(df.book60)/nrow(df.book)
```

```
## [1] 0.2030651
```

ii. Proportion of **DVD** sales transactions that are from the Web

```
# ii
df.dvd <- ST %>% filter(Product =="DVD")
df.dvdweb <- df.dvd %>% filter(Source=="Web")
nrow(df.dvdweb)/nrow(df.dvd)
```

```
## [1] 0.7630332
```

```
pnorm(25,mean=20, sd=4, lower.tail = FALSE)
```

```
## [1] 0.1056498
```