Sales Transactions Analysis

MinhChau

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
## Loading required package: carData
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve
3WBa
```

```
##
## Attaching package: 'rstatix'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following object is masked from 'package:rcompanion':
##
##     phi
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ lubridate 1.9.4     ✔ tibble    3.2.1
## ✔ purrr     1.0.4     ✔ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ psych::%+%()    masks ggplot2::%+%()
## ✖ psych::alpha()  masks ggplot2::alpha()
## ✖ dplyr::filter() masks rstatix::filter(), stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ purrr::lift()   masks caret::lift()
## ✖ dplyr::recode() masks car::recode()
## ✖ purrr::some()   masks car::some()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
ts to become errors
```

- 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.

```
#put in your working directory folder pathname ()
#import excel file into RStudio
ST <- read_excel("/Users/minhchau/Downloads/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
## # ℹ 1 more variable: `Time Of Day` <dttm>
```

1. Frequency distribution of Customer Profiles

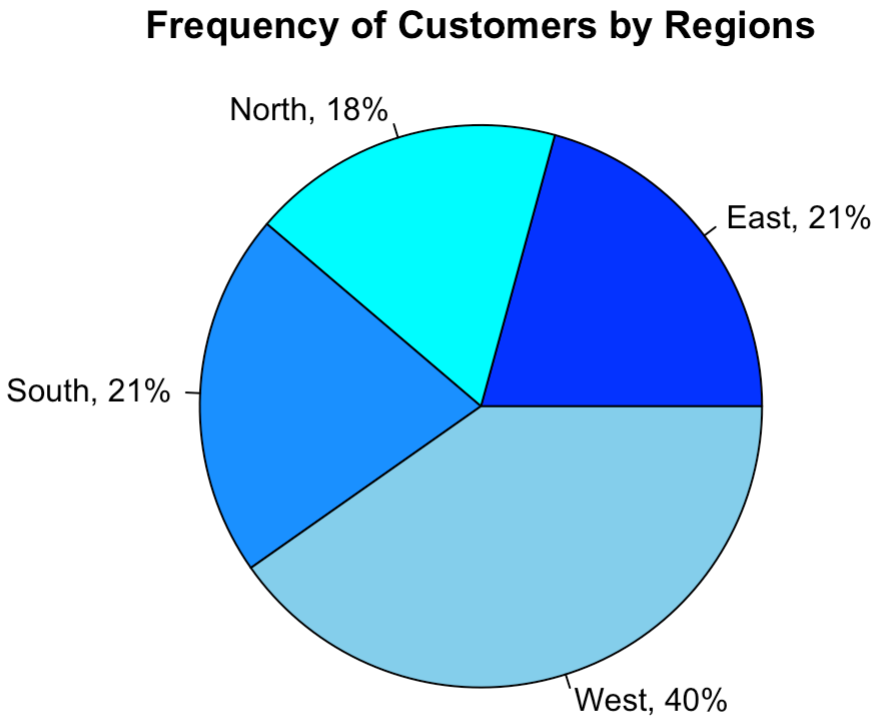
The manager would like to have a better understanding of the customer profiles. He would like the customer dashboard to display the following:

- 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.

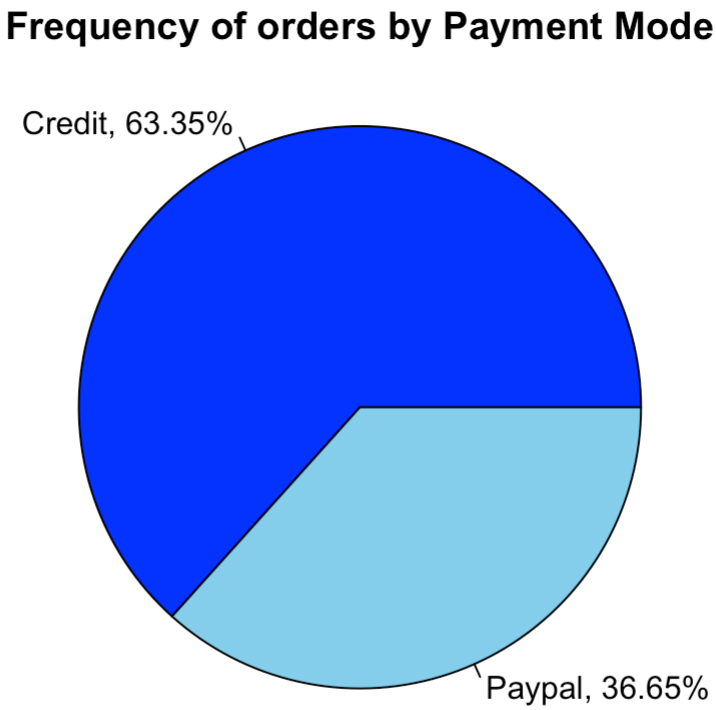
Frequency of Customers by Region

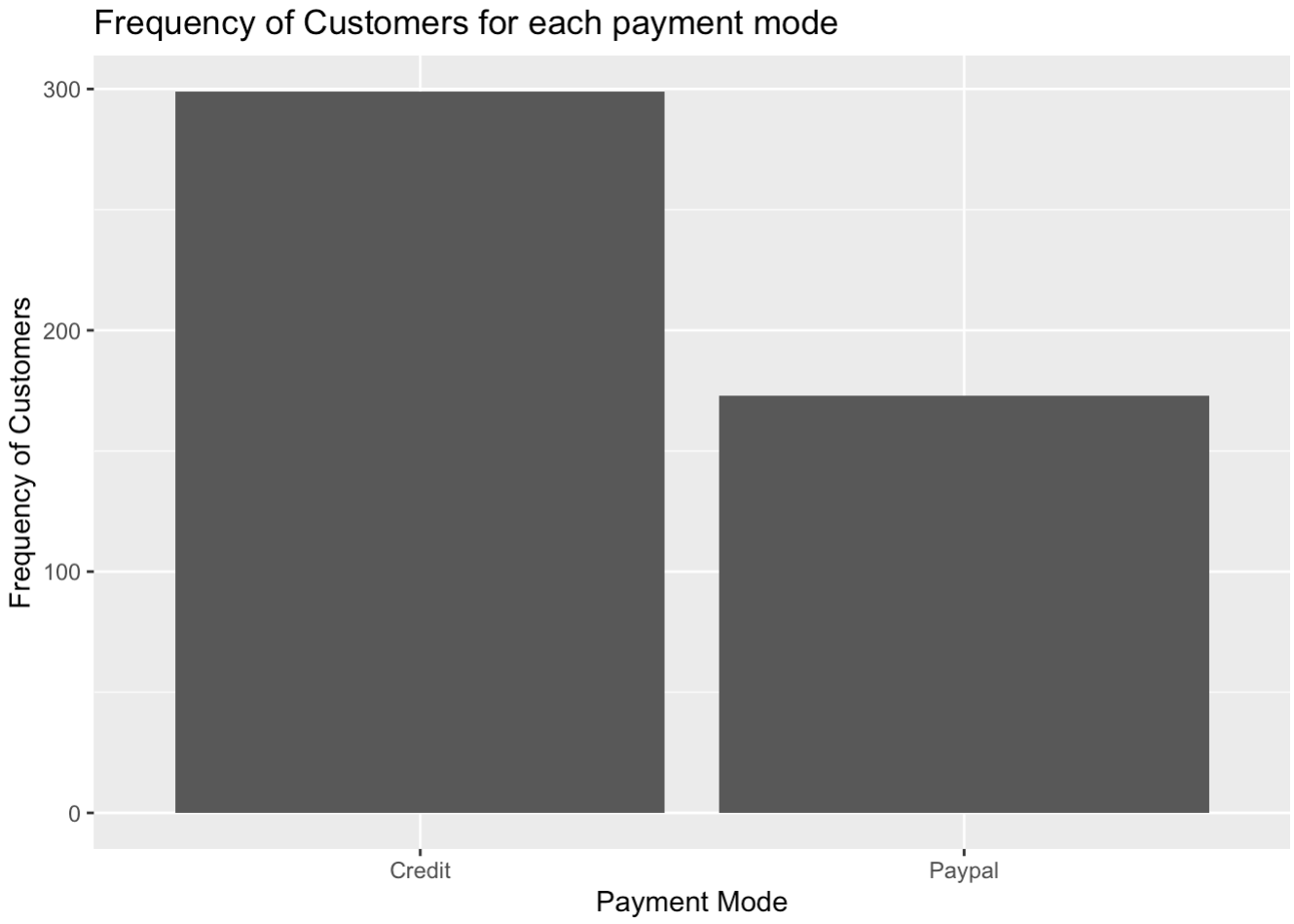
Region	n
East	98
North	85
South	99
West	190

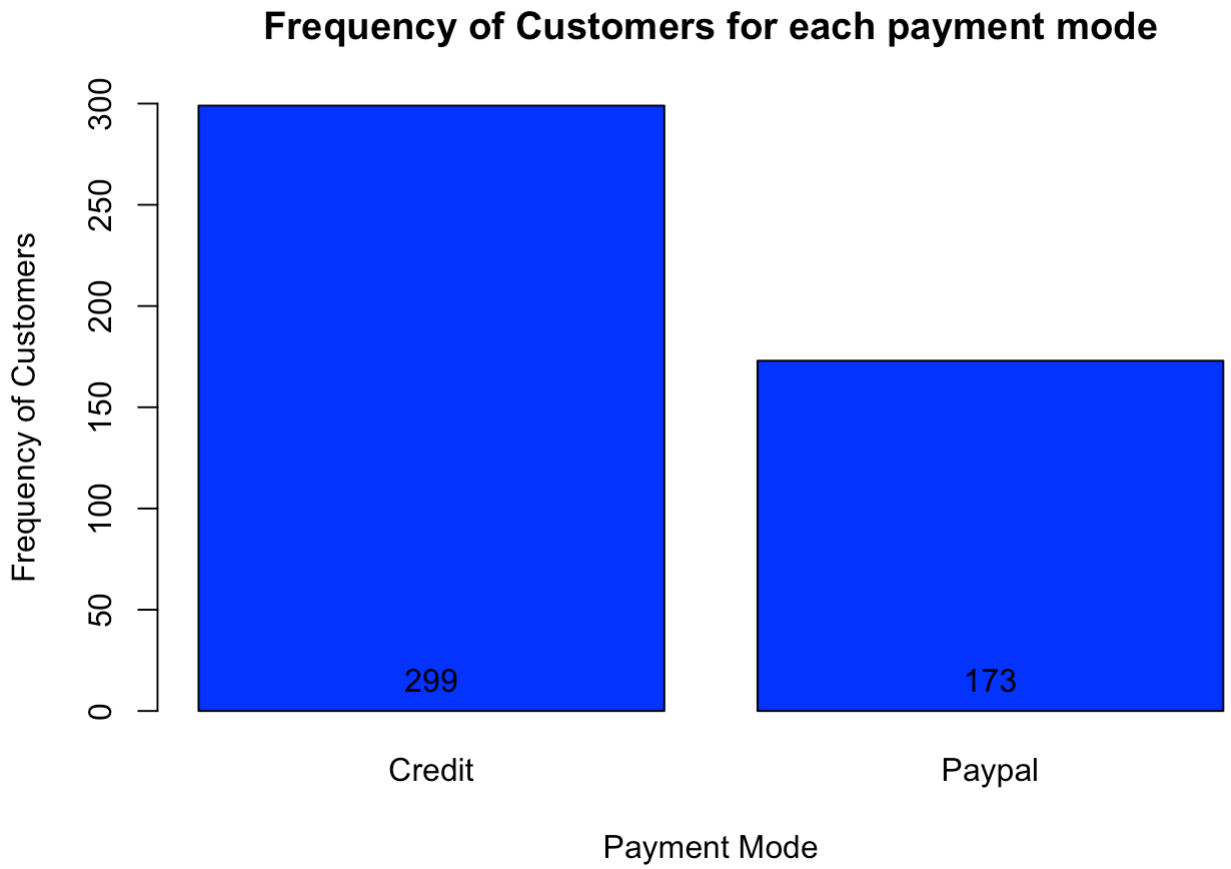


Frequency of Customers for each payment mode

Payment	n
Credit	299
Paypal	173


Frequency of orders by Payment Mode
Credit, 63.35%
Paypal, 36.65%


Frequency of Customers for each payment mode
300
200
100
0
Frequency of Customers
Credit
Paypal
Payment Mode



2. 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 or keep them? How would you handle these outliers?

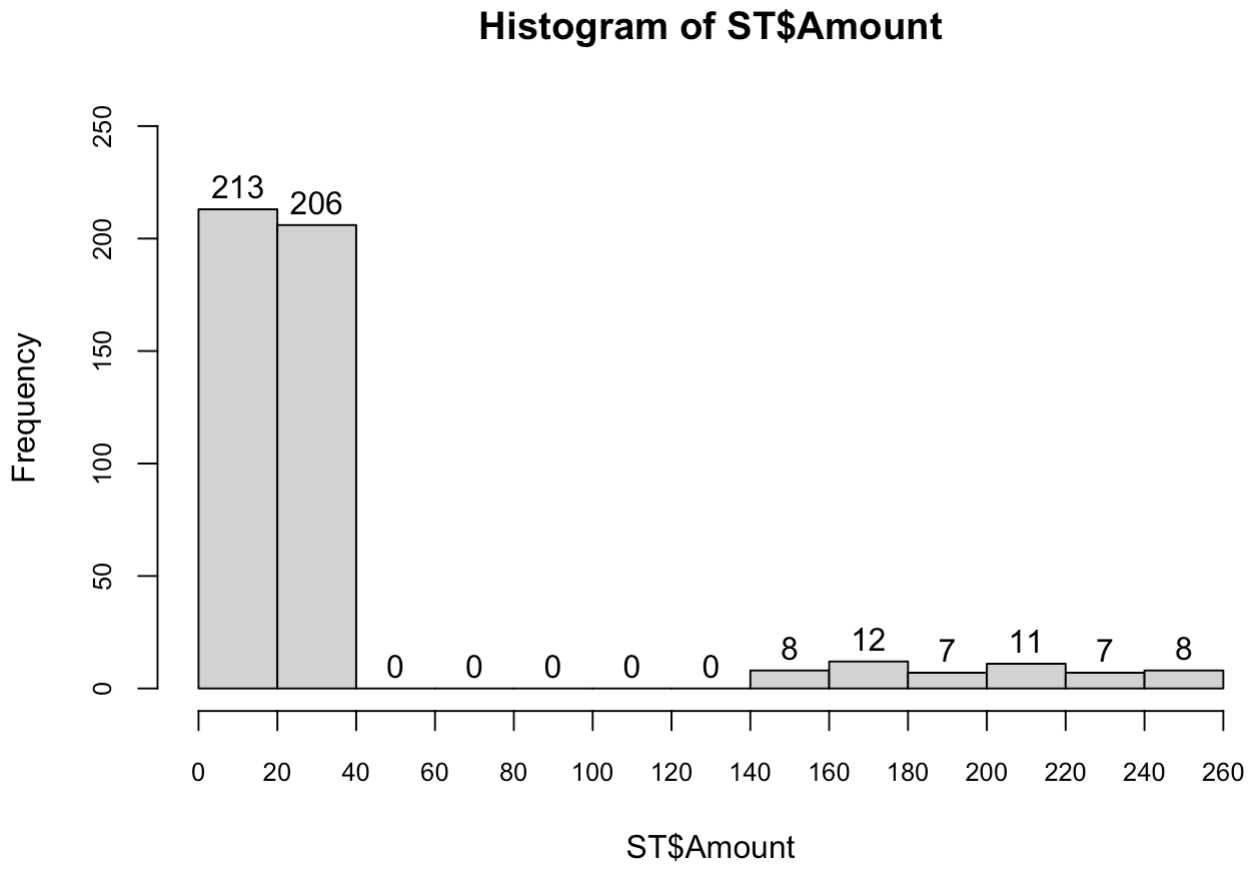
CODE

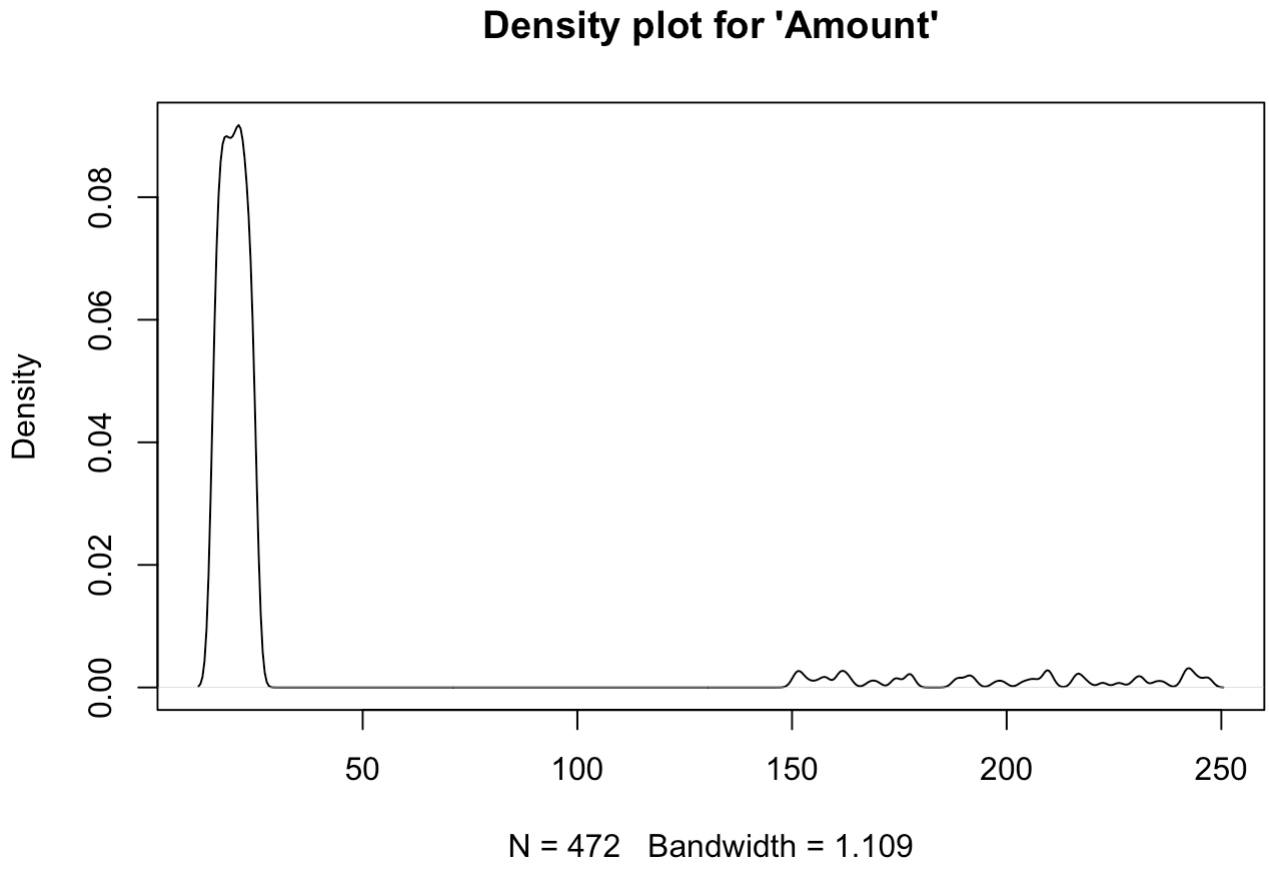
Descriptive Statistics for Amount

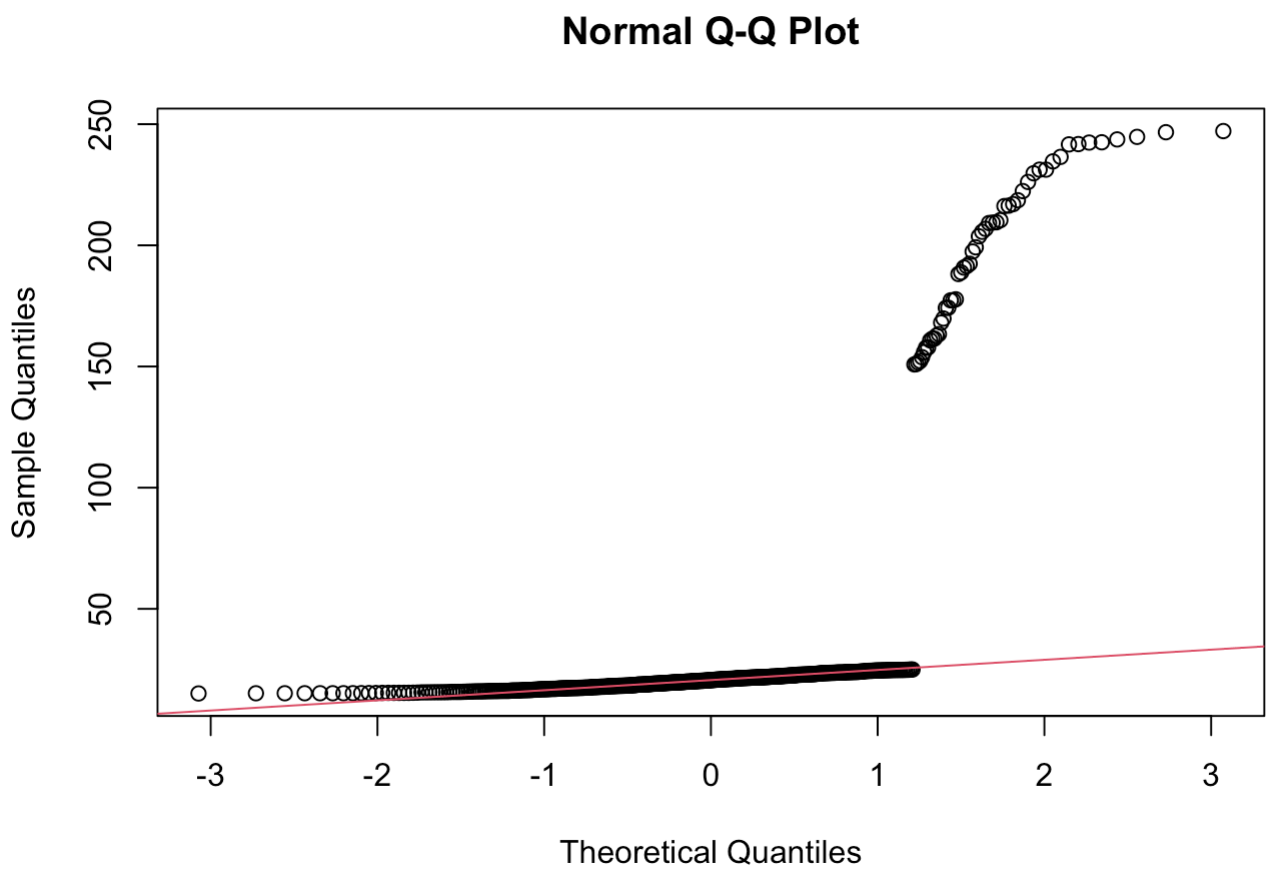
vars	n	mean	sd	median	skew	kurtosis
1	472	39.94581	57.32009	20.605	2.596053	5.080512

Description Statistics for Amount

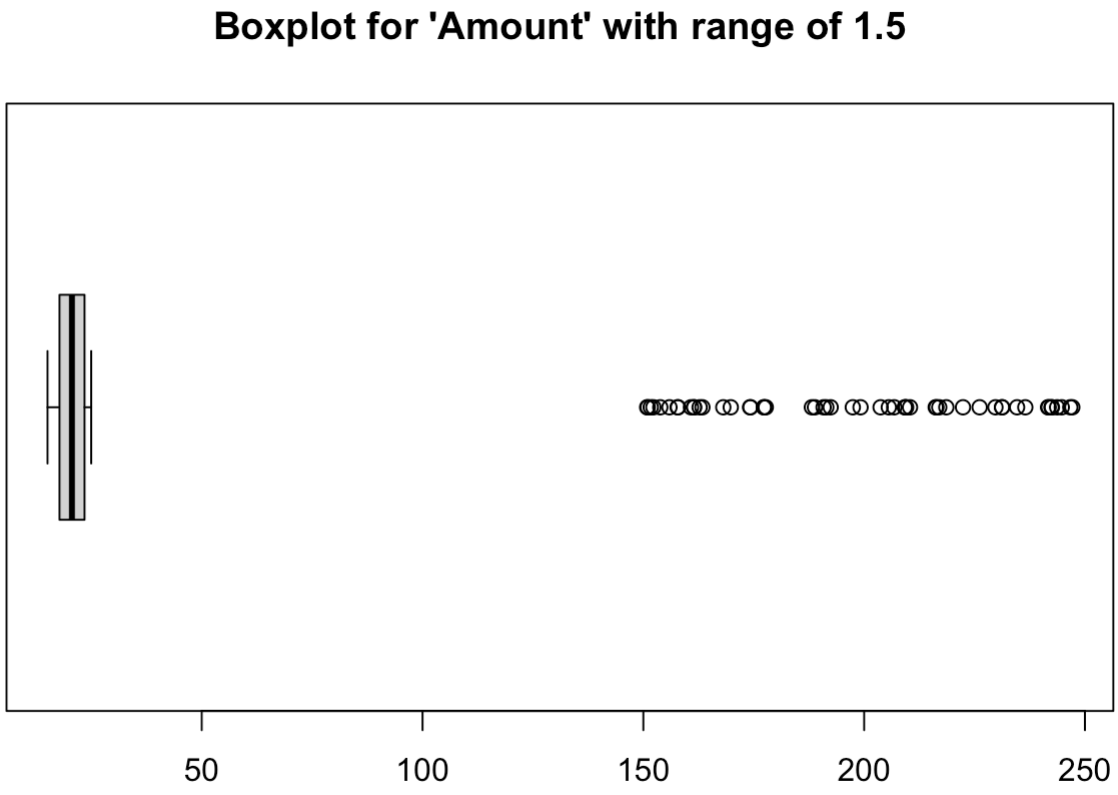
vars	n	mean	sd	median	skew	kurtosis
Amount	472	39.95	57.32	20.6	2.6	5.08

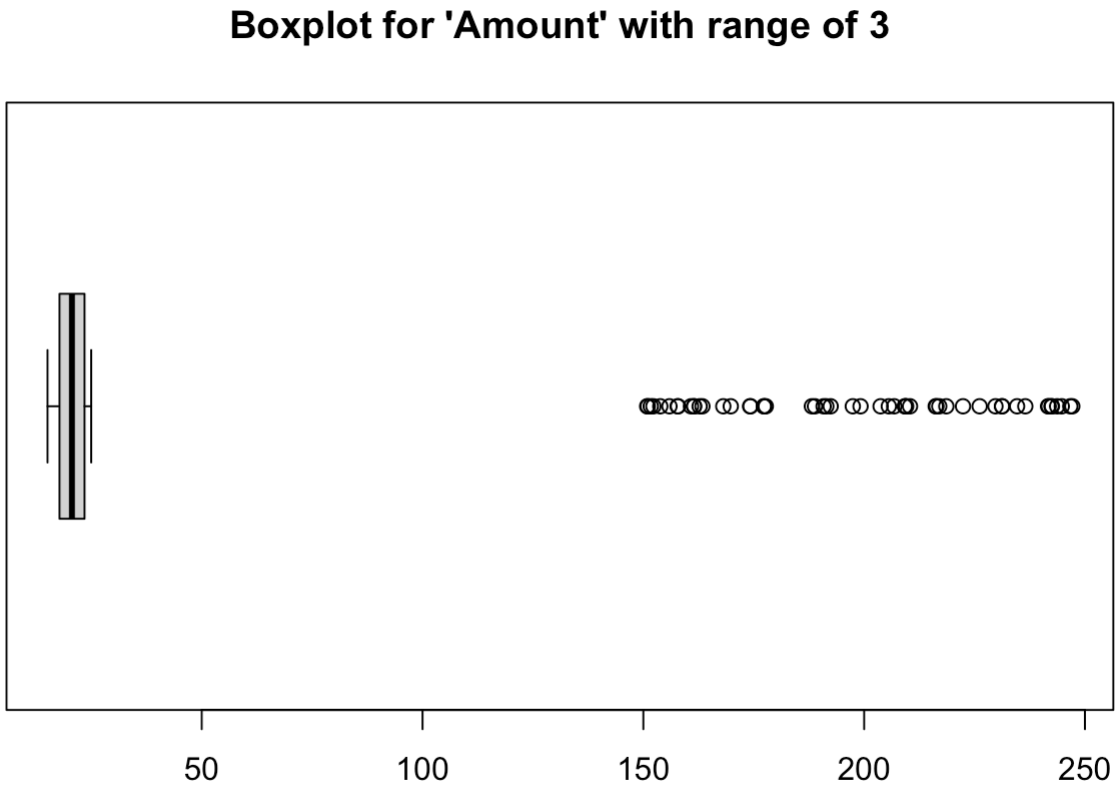

Histogram of ST$Amount
Frequency
0
50
100
150
200
250
213
206
0
0
0
0
0
8
12
7
11
7
8
0
20
40
60
80
100
120
140
160
180
200
220
240
260
ST$Amount


Density plot for 'Amount'
Density
0.00
0.02
0.04
0.06
0.08
50
100
150
200
250
N = 472 Bandwidth = 1.109



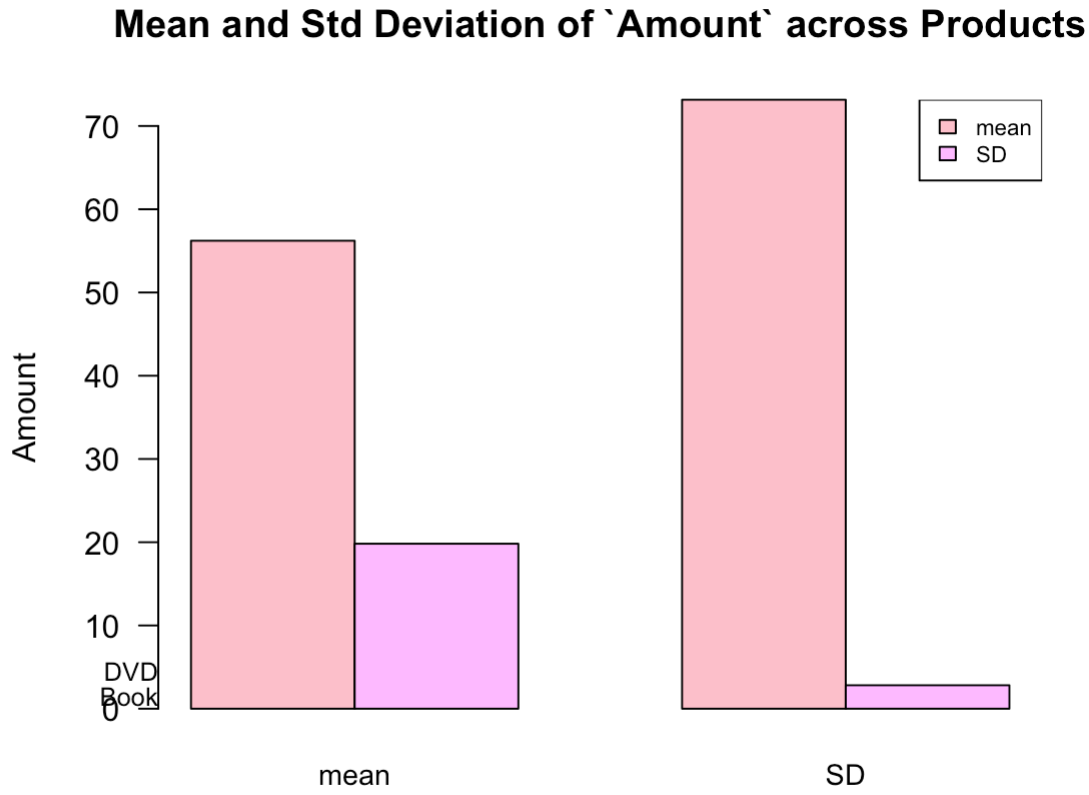

```
##
##  Shapiro-Wilk normality test
##
## data:  ST$Amount
## W = 0.42617, p-value < 2.2e-16
```

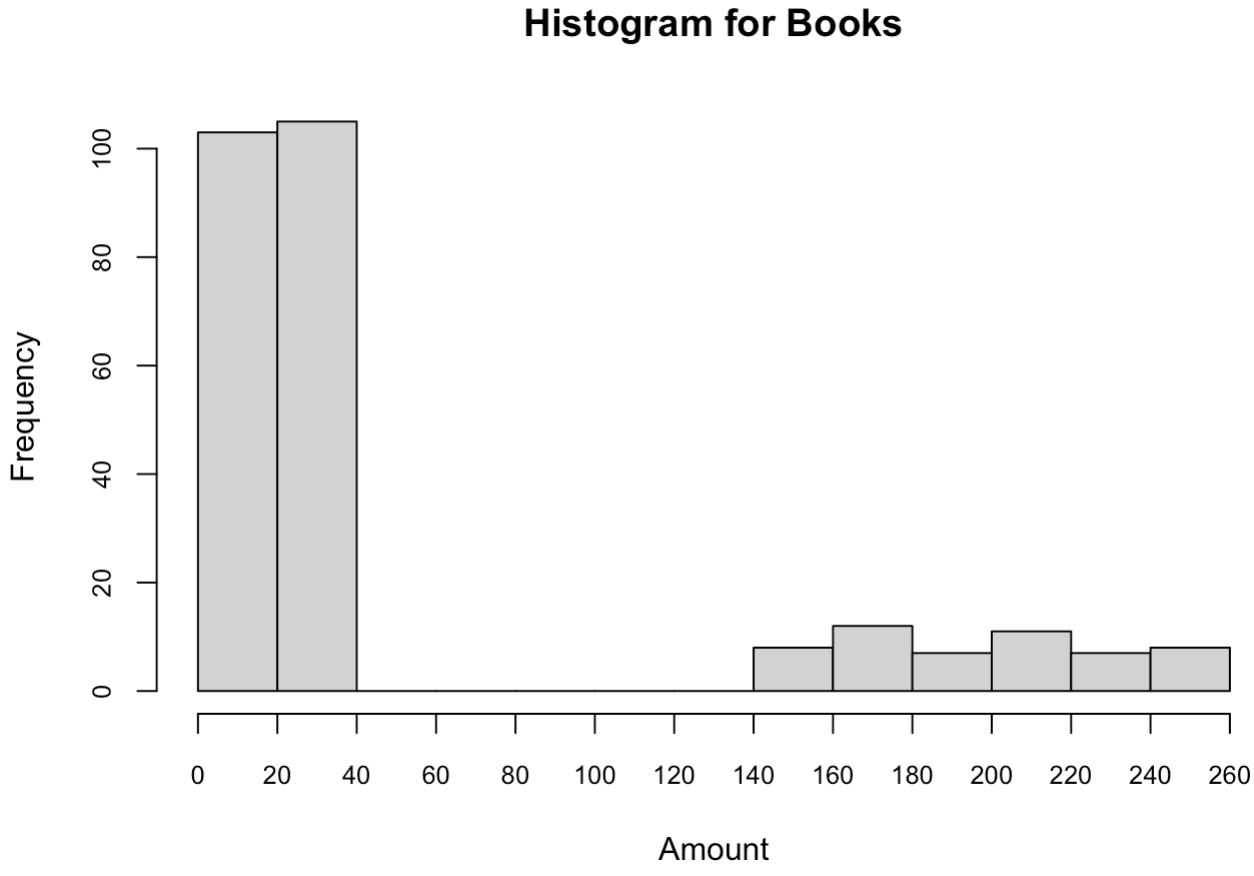


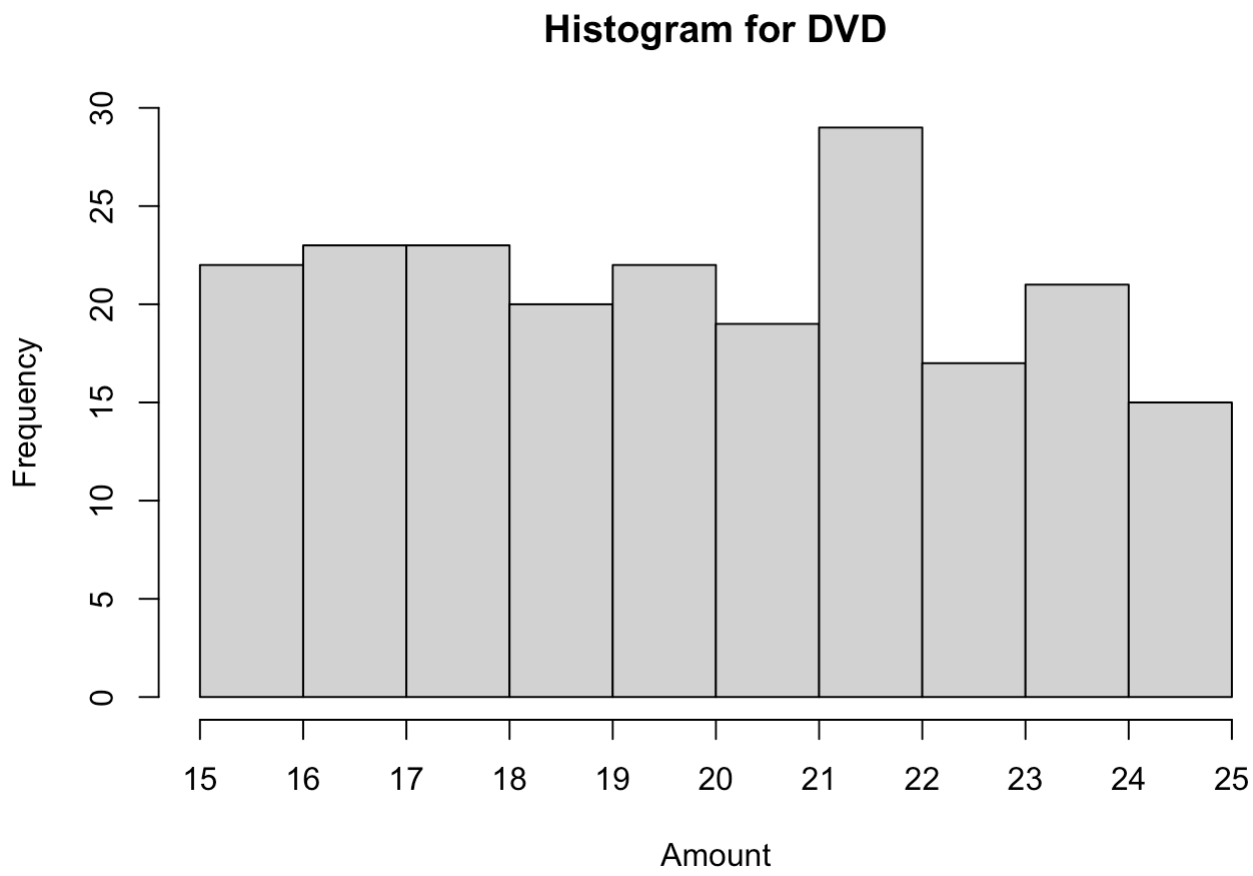


Product	mean	SD
Book	56.21559	73.15149

Product	mean	SD
DVD	19.82062	2.81961


Mean and Std Deviation of `Amount` across Products
mean
SD
Amount
70
60
50
40
30
20
10
0
DVD
Book
mean
SD


Histogram for Books
Frequency
0
20
40
60
80
100
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Amount



From the histograms, 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, we may wish to find out more before just discarding this data. • A discussion with the bookstore manager reveals that higher sales amount is due to the sales of rare/collector item books that tend to cost more. It might be interesting to examine if there is any difference between normal vs rare/collector item books. • Hence to deal with "outlier" here, one way is to analyse normal books and rare/collector books separately (we could create another variable, type, to indicate the type of books). This is something that needs to be discussed with the manager.

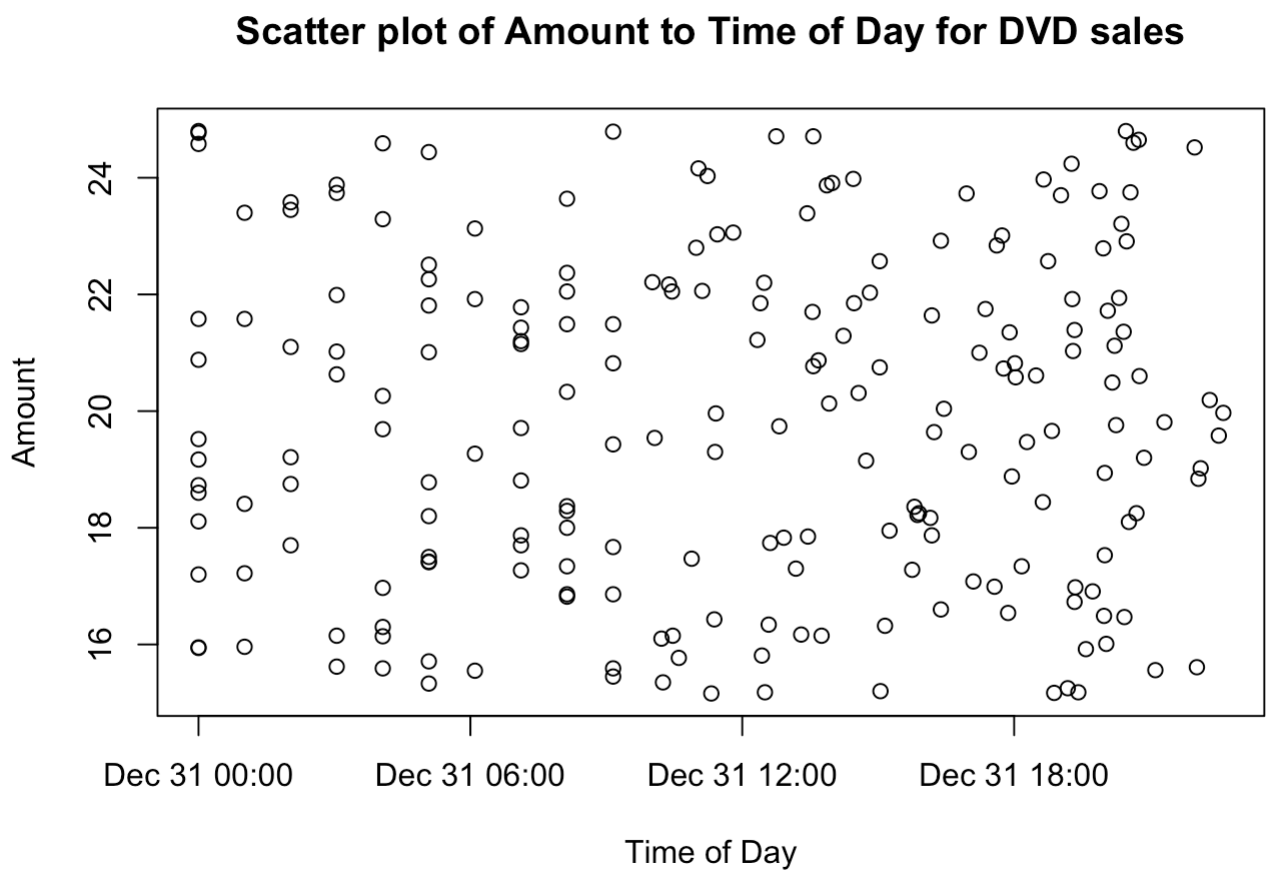
3. Checking Correlation

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.

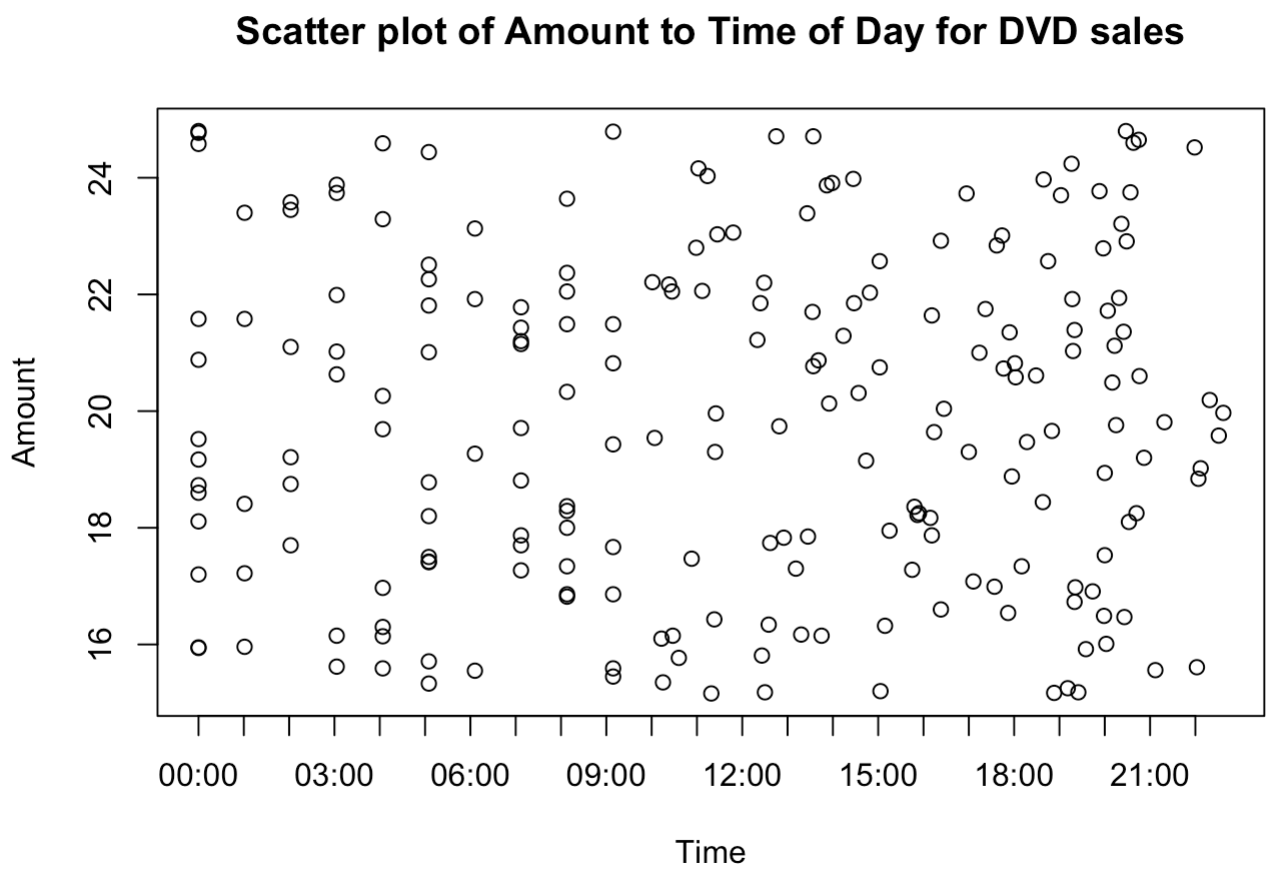
CODE

```
#plot
##(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")
```




```
#Use axis.POSIXct to reformat the x axis to keep only the hours and mins.
dvddata <- ST %>% filter(Product == "DVD")
time <- dvddata$`Time Of Day` # you could use the original variable too
plot(time, dvddata$Amount, xlab = "Time", ylab = "Amount", main = "Scatter plot of Am
ount to Time of Day for DVD sales",
     xaxt = "n") # Suppress the x-axis

# Add custom x-axis labels with only the time
axis.POSIXct(1, at = seq(min(time), max(time), by = "hour"),
             labels = format(seq(min(time), max(time), by = "hour"), "%H:%M"))
```




```
#Stat measurement
cor(as.numeric(tab.DVD$`Time Of Day`), tab.DVD$Amount)
```

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

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

```
##
##  Pearson's product-moment correlation
##
## data:  as.numeric(tab.DVD$`Time Of Day`) and tab.DVD$Amount
## t = 0.46122, df = 209, p-value = 0.6451
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1036276  0.1662399
## sample estimates:
##       cor
## 0.03188728
```

4. 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 greater than $25?

Please type your answer below.

CODE

```
# i. Proportion of Book sales transactions that have Amount greater than $60
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
df.dvd <- ST %>% filter(Product =="DVD")
df.dvdweb <- df.dvd %>% filter(Source=="Web")
nrow(df.dvdweb)/nrow(df.dvd)
```

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

```
# The last line in the image seems to be unrelated to the proportions.
# It calculates the probability of a value being greater than 25 in a normal distribu
tion with mean 20 and standard deviation 4.
pnorm(25, mean=20, sd=4, lower.tail = FALSE)
```

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

5. Computing Interval Estimates

- i. compute the 99% for the mean of `Amount` for DVD sale transactions. Could the company conclude with 99% confidence level that the true mean `Amount` for DVD sale transactions is not equal to $20?
- ii. compute the 90% confidence interval for proportion of DVD sale transactions with sales amount being greater than $22. Explain to the store manager what this confidence interval means.
- iii. compute the 95% prediction interval for `Amount` for sales of DVD. Explain to the store manager what this prediction interval mean?

CODE

```
#possibility not 20, confidence interval, cant reject

dfd <- ST %>% filter(Product == "DVD")
uciatm99  <- mean(dfd$Amount) - qt(0.005, df=nrow(dfd) - 1)*sd(dfd$Amount)/sqrt(nrow
(dfd))
lciatm99  <- mean(dfd$Amount) + qt(0.005, df=nrow(dfd) - 1)*sd(dfd$Amount)/sqrt(nrow
(dfd))
 print(cbind(uciatm99, lciatm99), digits = 4)
```

```
##      uciatm99 lciatm99
## [1,]    20.33    19.32
```

It is perhaps easier to understand this problem, using our intuition regarding Hypothesis Testing. Null Hypothesis H0: True population mean Amount for DVD sales transactions = $20. Alternat Hypothesis H1: True population mean Amount for DVD sales transactions != $20 (two-tailed test)

1st method : - Generate 99% confidence interval: (19.3, 20.3) - Conclusion: Since $20 lies within the 99% confidence interval, we have insufficient evidence at the 99% level of confidence to reject the null hypothesis that True population mean Amount for DVD sales transactions = $20. - Note: DO NOT MENTION THAT “we accept the Null Hypothesis”. - We never accept the Null Hypothesis, we only: (a) reject Null Hypothesis, or (b) fail to reject Null Hypothesis.

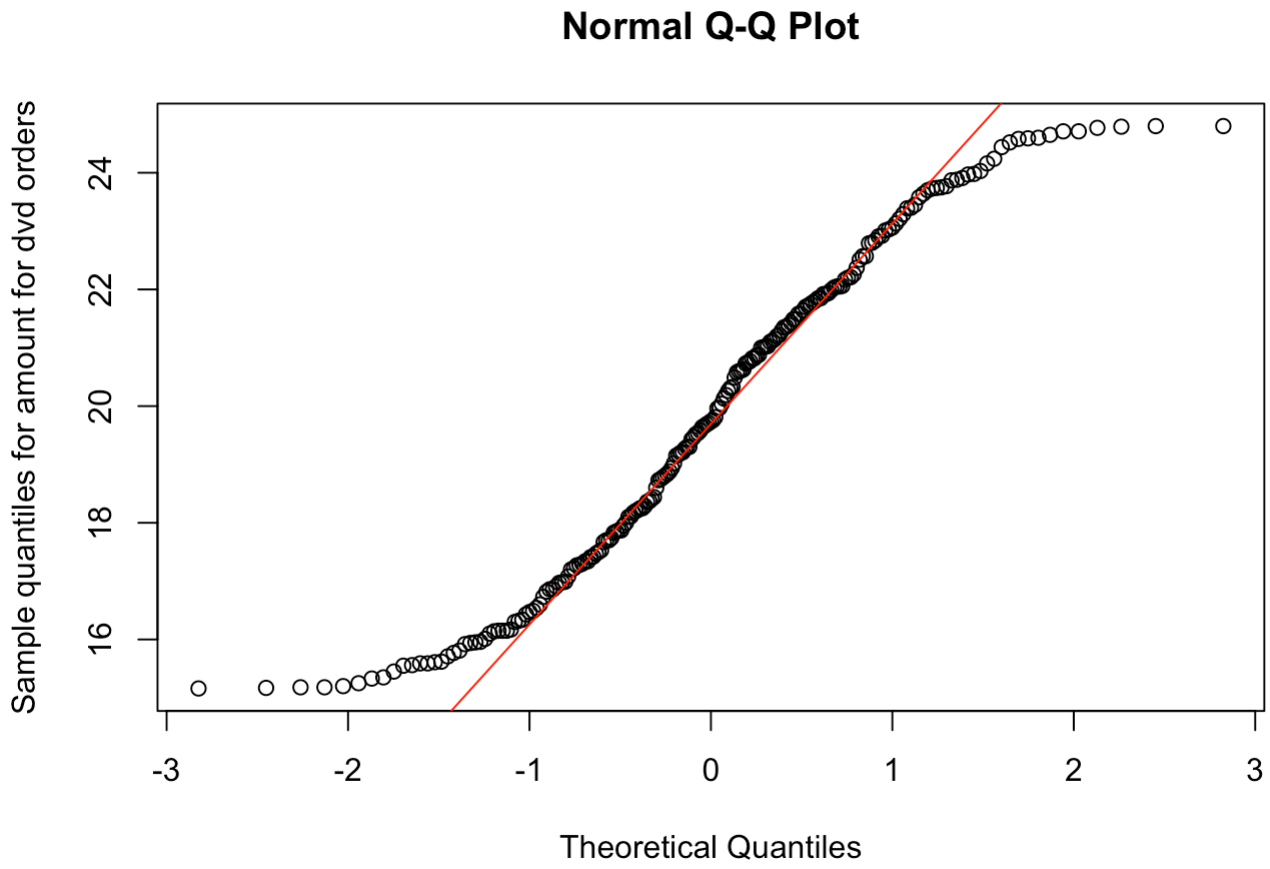
2nd method : - Use t-test - The output will include a p-value, which you compare to your significance level (α = 0.01, since we’re using a 99% confidence level here). - Decision rule: - If p-value ≤ α (0.01): We reject the null hypothesis. This means there is evidence that the population mean is significantly different from $20. - If p-value > α (0.01): We fail to reject the null hypothesis. This means there is insufficient evidence to conclude that the mean is different from $20.

```
df22 <- ST %>% filter(Amount > 22)
pd22 <- nrow(df22)/nrow(dfd)
lcipd22 <- pd22 + (qnorm(0.05)*sqrt(pd22*(1-pd22)/nrow(dfd)))
ucipd22 <- pd22 - (qnorm(0.05)*sqrt(pd22*(1-pd22)/nrow(dfd)))
print(cbind(lcipd22, ucipd22), digits=3)
```

```
##      lcipd22 ucipd22
## [1,]   0.761    0.85
```

The 90% confidence interval for the proportion of DVD sale transactions with an amount greater than $22 is (0.202, 0.3). This means that we are 90% confident that the true proportion of DVD transactions with sales greater than $22 lies between 0.202 and 0.3. In other words, if we were to repeatedly take samples from all DVD sales and compute the proportion, 90% of the intervals we calculate would contain the true proportion.

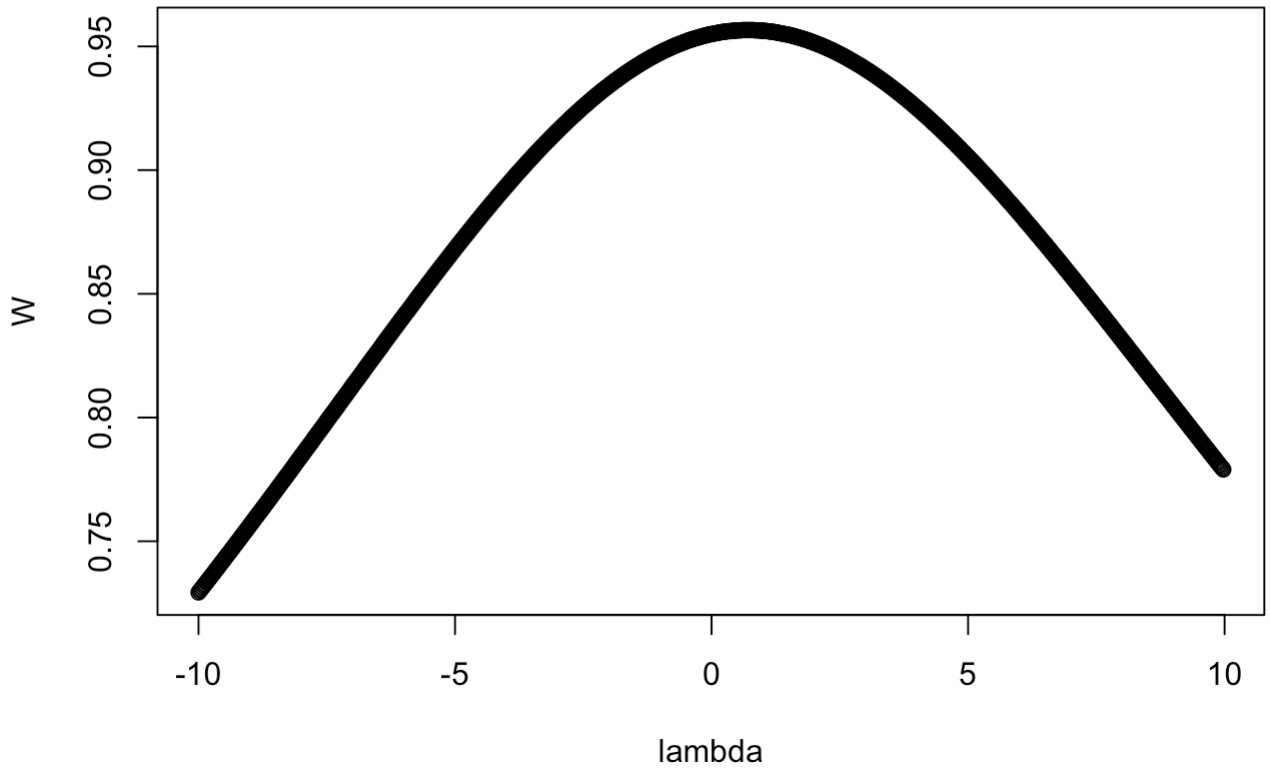
```
# III. compute the 95% prediction interval for `Amount` for sales of DVD. Explain to
the store manager what this prediction interval mean?
amount_clean <- na.omit(dfd$Amount)
amount_clean <- as.numeric(amount_clean)
qqnorm(amount_clean,
       ylab = "Sample quantiles for amount for dvd orders")
qqline(amount_clean, col="red")
```



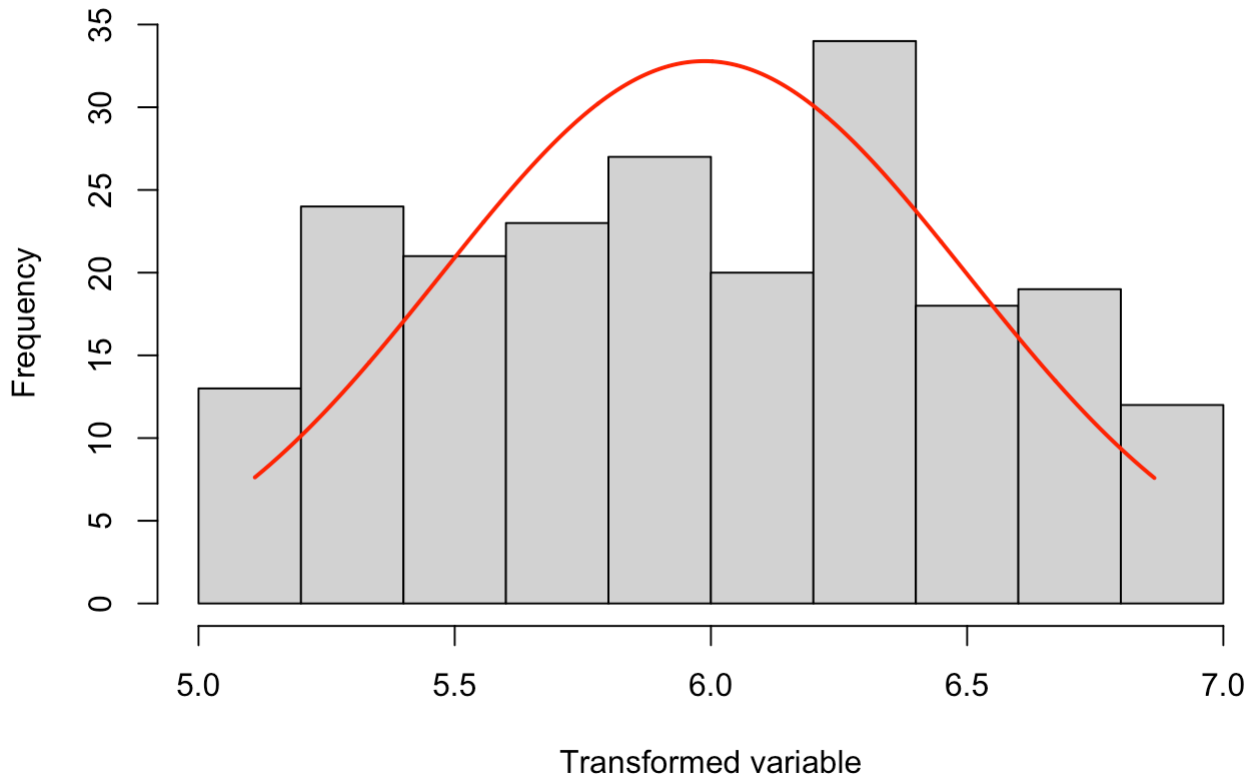

```
shapiro.test(amount_clean)
```

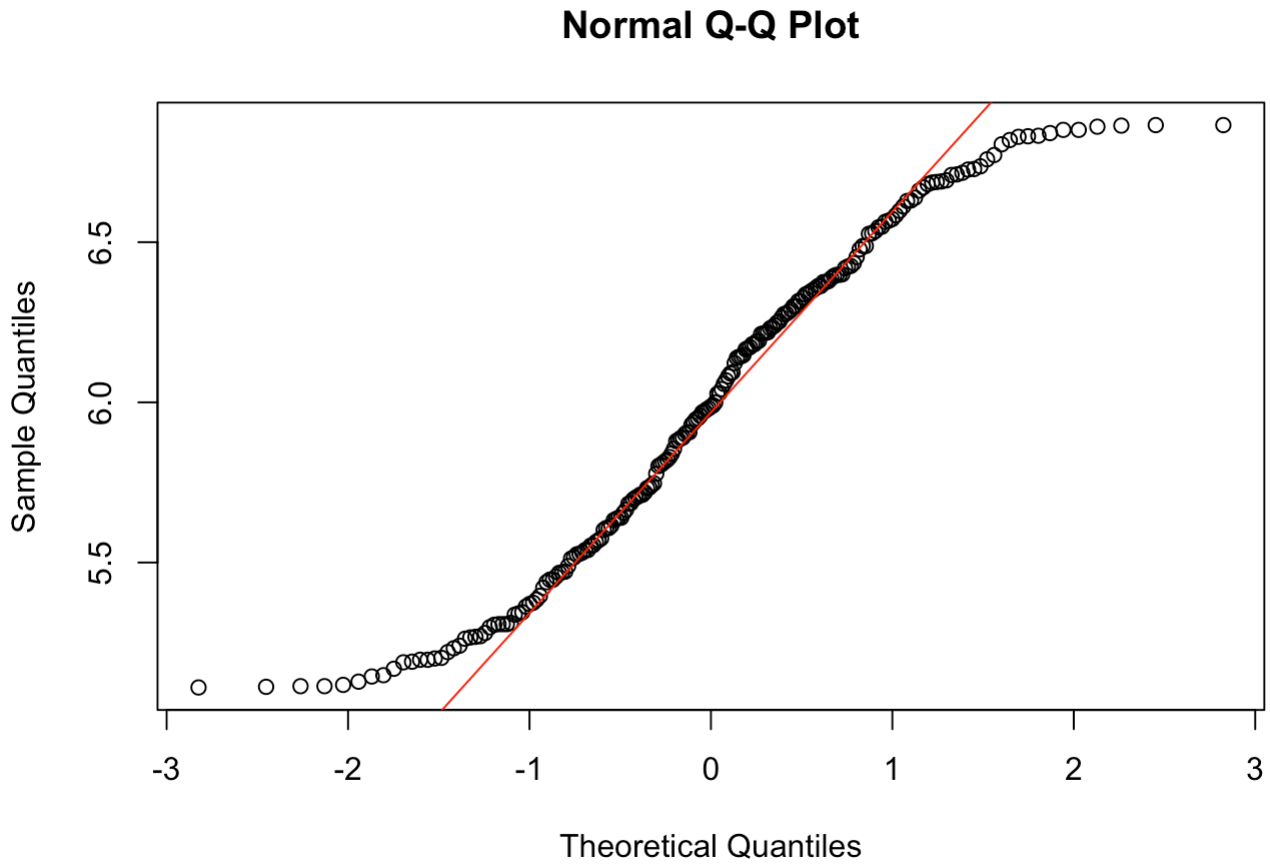
```
##
##  Shapiro-Wilk normality test
##
## data:  amount_clean
## W = 0.95635, p-value = 4.703e-06
```

```
dfd$Amt.t = transformTukey(amount_clean, plotit=TRUE)
```




```
##
##     lambda      W Shapiro.p.value
## 425    0.6 0.9566       4.959e-06
##
## if (lambda >  0){TRANS = x ^ lambda}
## if (lambda == 0){TRANS = log(x)}
## if (lambda <  0){TRANS = -1 * x ^ lambda}
```


Frequency
0
5
10
15
20
25
30
35
5.0
5.5
6.0
6.5
7.0
Transformed variable


Normal Q-Q Plot
Sample Quantiles
5.5
6.0
6.5
-3
-2
-1
0
1
2
3
Theoretical Quantiles

```
mnamt <- mean(dfd$Amount)
sdamt <- sd(dfd$Amount)
lpi.amt <- mnamt + (qt(0.025, df = (nrow(dfd)-1))*sdamt*sqrt(1+1/nrow(dfd)))
upi.amt <- mnamt - (qt(0.025, df = (nrow(dfd)-1))*sdamt*sqrt(1+1/nrow(dfd)))
cbind(lpi.amt, upi.amt)
```

```
##       lpi.amt  upi.amt
## [1,] 14.24909 25.39214
```

```
#The 95% prediction interval for DVD sale amounts is (14.25, 25.39). This means that for a single future DVD sale, we can be 95% confident that the sale amount will fall within this range.
```

Q1.(b) Hypothesis Testing

The store manager would like to draw some conclusions from the sample sales transaction data. He would like to retain all the data for the analyses. Please help him to set up and test the following hypotheses.You may assume that `Amount` is normally distributed here

- i. The proportion of book sales transactions with `Amount` greater than $50 is at least 25 percent of book sales transactions.
- ii. The mean sales amount for books is the same as for dvds.
- iii. The mean sales amount for CollectorBook is greater than mean sales amount for Book (ie normal or non-collector books). You may use the definition from T4 Part 1 Q1biii where the outliers identified from the boxplot of range 3 is used to indicate CollectorBook.
- iv. The mean sales amount for dvds is the same across all 4 regions.

CODE

```
#just use p value
#one sample test for proportion => use z statistic => calculate the proportion with Amount greater than $50 is at least 25 percent of book sales transactions.

book <- ST %>%  filter(Product == "Book")
bk50 <- book %>% filter(Amount > 50)
pbk50 <- nrow(bk50)/nrow(book)

z<- (pbk50-0.25)/sqrt(0.25*(1-0.25)/nrow(book))
z
```

```
## [1] -1.751117
```

```
cv95<- qnorm(0.05)
cv95
```

```
## [1] -1.644854
```

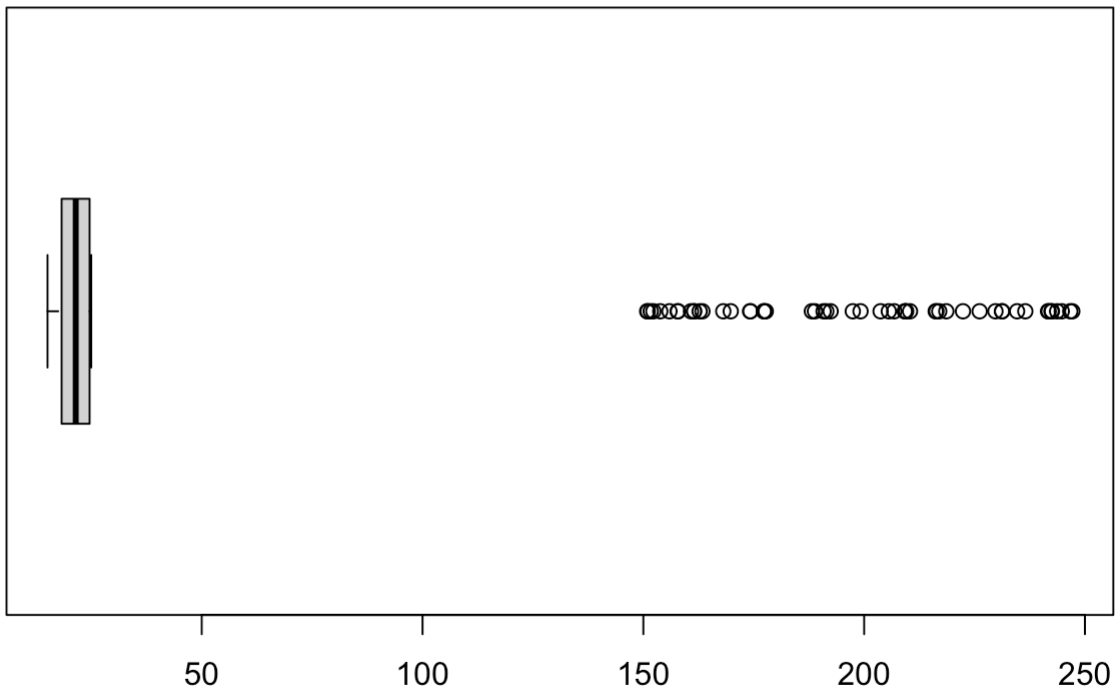
```
z<- cv95
pnorm(z)
```

```
## [1] 0.05
```

```
#II. The mean sales amount for books is the same as for dvds
ST$Amount <- as.numeric(as.character(ST$Amount))
t.test(Amount ~ Product, data=ST)
```

```
##
##  Welch Two Sample t-test
##
## data:  Amount by Product
## t = 8.0304, df = 260.96, p-value = 3.344e-14
## alternative hypothesis: true difference in means between group Book and group DVD
is not equal to 0
## 95 percent confidence interval:
##  27.47079 45.31916
## sample estimates:
## mean in group Book  mean in group DVD
##           56.21559           19.82062
```

```
#III. The mean sales amount for CollectorBook is greater than mean sales amount for B
ook (ie normal or non-collector books). You may use the definition from T4 Part 1 Q1b
iii where the outliers identified from the boxplot of range 3 is used to indicate Col
lectorBook.
#book sale amount sample of hypothetical, based on existing => make conclusion about
all transaction ever make
book$Amount <- as.numeric(as.character(book$Amount))
boxplot.bk <- boxplot(book$Amount, horizontal = TRUE, range = 3)
```




```
book1 <- book %>% mutate(Pdt_type = ifelse(
  Amount %in% boxplot.bk$out, "CollectionBook", "Book"
))
t.test(book1$Amount~book1$Pdt_type, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  book1$Amount by book1$Pdt_type
## t = -40.548, df = 52.22, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Book and group Coll
ectionBook is less than 0
## 95 percent confidence interval:
##       -Inf -170.5442
## sample estimates:
##           mean in group Book mean in group CollectionBook
##                     20.09216                    197.98302
```

```
#IV. The mean sales amount for dvds is the same across all 4 regions.
ST.dvd <- ST %>% filter(Product == "DVD")
table(ST.dvd$Region)
```

```
##
##  East North South  West
##    42    42    37    90
```

```
bartlett.test(Amount~Region, ST.dvd)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Amount by Region
## Bartlett's K-squared = 1.3863, df = 3, p-value = 0.7087
```