Descriptive Analytics through Tabulation, Graphs & Statistical Measures Workshop

TBA2102 2020/2021 Semester 2
Tutorial 4

STRUCTURE OF TUTORIALS

Duration:

45 mins

Content:

- Cover previous week’s tutorial assignment
- Tutorial 4

Tutorial 3 Assignment

PURCHASE ORDER

The `Purchase Orders.xlsx` data set contains data on all items that an aircraft component manufacturing company has purchased over the past 3 months. Each of the column is defined as follows:

- `Supplier`	Supplier of items purchased
- `Order No.`	Order Number of the items purchased
- `Item No.`	A categorical variable used to identify the item
- `Item Description`	Description of the item purchased
- `Item Cost`	Item unit cost
- `Quantity`	Number of items bought in the purchase order
- `Cost per order`	Total cost of the order
- `A/P Terms (Months)`	Suppliers' Accounts Payable (A/P) terms
- `Order Date`	Items order date
- `Arrival Date`	Items arrival date

LET'S TAKE A LOOK AT THE DATA

```
glimpse(PO)
```

- glimpse is from dplyr package
- An alternative to str()

```
Rows: 94
Columns: 11
$ Supplier              <chr> "Hulkey Fasteners", "Alum Sheeting", "Fast-Tie Aerospace"...
$ `Order No.`           <chr> "Aug11001", "Aug11002", "Aug11003", "Aug11004", "Aug11005...
$ `Item No.`            <dbl> 1122, 1243, 5462, 5462, 5319, 5462, 4312, 7258, 6321, 546...
$ `Item Description`    <chr> "Airframe fasteners", "Airframe fasteners", "Shielded Cab...
$ `Item Cost`           <dbl> 4.25, 4.25, 1.05, 1.05, 1.10, 1.05, 3.75, 90.00, 2.45, 1....
$ Quantity              <dbl> 19500, 10000, 23000, 21500, 17500, 22500, 4250, 100, 1300...
$ `Cost per order`      <dbl> 82875.00, 42500.00, 24150.00, 22575.00, 19250.00, 23625.0...
$ `A/P Terms (Months)`  <dbl> 30, 30, 30, 30, 30, 30, 30, 45, 30, 30, 30, 30, 30, 30, 3...
$ `Order Date`          <dttm> 2011-08-05, 2011-08-08, 2011-08-10, 2011-08-15, 2011-08-...
$ `Arrival Date`        <dttm> 2011-08-13, 2011-08-14, 2011-08-15, 2011-08-22, 2011-08-...
$ `Arrival Time`        <dbl> 8, 6, 5, 7, 11, 6, 7, 3, 10, 8, 11, 4, 6, 8, 9, 9, 5, 9, ...
```

Other functions that can help you explore the data

- View(PO)
- str(PO)
- head(PO)
- lapply(PO,class) --- check the data type of all the variables

QUESTION 2A: PURCHASE ORDER DASHBOARD

The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier**
- **Item Description**
- **Cost per order**
- **Arrival Time**

i. Arrival Time is the difference between Arrival Date and Order Date. Create this variable in the dataframe. (1 mark)
ii. Generate the charts and tables to view the frequency distributions of the 4 variables of interest to the manager. (4 marks: 0.5 marks per table/chart)
iii. Generate the appropriate chart to display the relationship between Cost per order and Arrival Time. (1 mark)
iv. Describe in your answer below your observations from each of the charts. (2.5 marks)

QUESTION 2A: PURCHASE ORDER DASHBOARD

The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier, Item Description, Cost per order and Arrival Time**

i. Arrival Time is the difference between Arrival Date and Order Date. Create this variable in the dataframe. (1 mark)

```
PO$`Arrival Time`<-as.numeric(PO$`Arrival Date`-PO$`Order Date`)
```

We could also use a function **mutate** from dplyr to create this variable in the dataframe.

QUESTION 2A: PURCHASE ORDER DASHBOARD

The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier, Item Description, Cost per order and Arrival Time**

ii. Generate the charts and tables to view the frequency distributions of the 4 variables of interest to the manager. (4 marks: 0.5 marks per table/chart)

Supplier

```
SuppFq<-PO%>%
  count(`Supplier`)
kable(SuppFq, caption = "Frequency Distribution for Supplier")
Suppbar <- SuppFq$n
```

Frequency Distribution for Supplier

Supplier	n
Alum Sheeting	8
Durrable Products	13
Fast-Tie Aerospace	15
Hulkey Fasteners	15
Manley Valve	11
Pylon Accessories	5
Spacetime Technologies	12
Steelpin Inc.	15

- Orders were most frequently from 3 Suppliers: Steelpin Inc, Hulkey Fasteners & Fast-Tie Aerospace)
- Orders were least frequently from Pylon Accessories.

QUESTION 2A: PURCHASE ORDER DASHBOARD

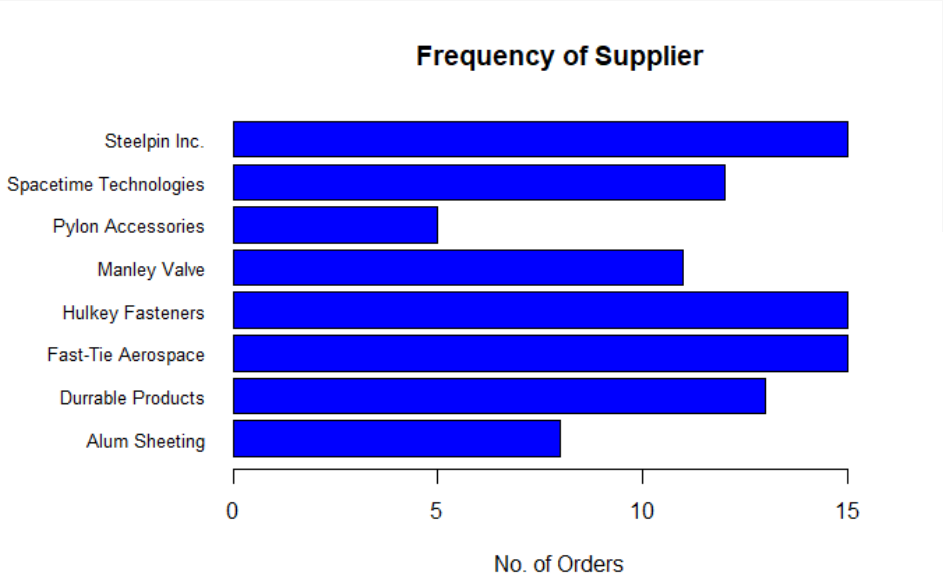
The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier, Item Description, Cost per order and Arrival Time**

ii. Generate the charts and tables to view the frequency distributions of the 4 variables of interest to the manager. (4 marks: 0.5 marks per table/chart)

Supplier

```
# Horizontal
# default is (5,4,4,2), I'm adding a bigger left margin for the barchart
par(mar=c(5,10,4,2))
barplot(Suppbar,names.arg=SuppFq$Supplier,
        col="blue", main="Frequency of Supplier",
        cex.names = 0.8,
        xlab="No. of Orders",
        xlim=c(0,16),
        horiz=TRUE,
        las=1)
```



QUESTION 2A: PURCHASE ORDER DASHBOARD

The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier, Item Description, Cost per order and Arrival Time**

ii. Generate the charts and tables to view the frequency distributions of the 4 variables of interest to the manager. (4 marks: 0.5 marks per table/chart)

Item Description

```
IdFq<-PO%>%
  count(`Item Description`)
kable(IdFq, caption = "Frequency Distribution for Item Description")
Idbar <- IdFq$n
```

Frequency Distribution for Item Description

Item Description	n
Airframe fasteners	14
Bolt-nut package	11
Control Panel	4
Door Decal	2
Electrical Connector	8
Gasket	10
Hatch Decal	2
Machined Valve	4
O-Ring	12
Panel Decal	1
Pressure Gauge	7
Shielded Cable/ft.	11
Side Panel	8

QUESTION 2A: PURCHASE ORDER DASHBOARD

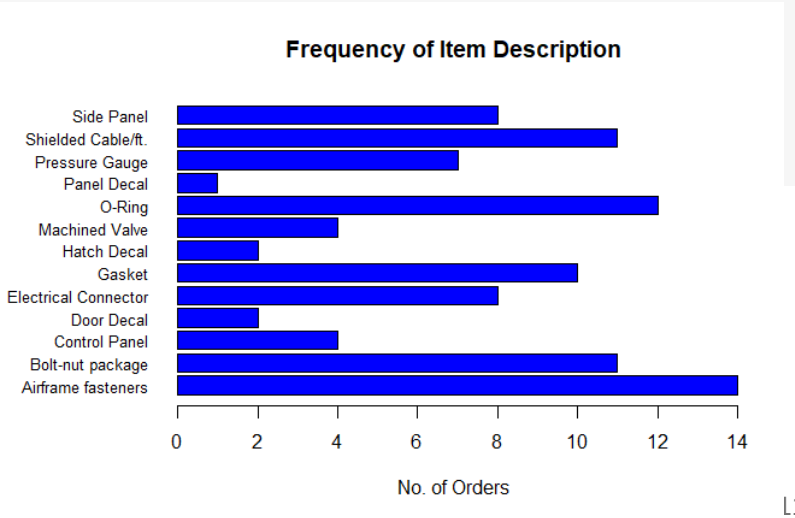
The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier, Item Description, Cost per order and Arrival Time**

ii. Generate the charts and tables to view the frequency distributions of the 4 variables of interest to the manager. (4 marks: 0.5 marks per table/chart)

Item Description

```
# Horizontal
# default is (5,4,4,2), I'm adding a bigger left margin for the barchart
par(mar=c(5,10,4,2))
barplot(Idbar,
        names.arg=IdFq$`Item Description`,
        col="blue", main="Frequency of Item Description",
        cex.names = 0.8,
        xlab="No. of Orders",
        horiz=TRUE,
        las=1)
```



QUESTION 2A: PURCHASE ORDER DASHBOARD

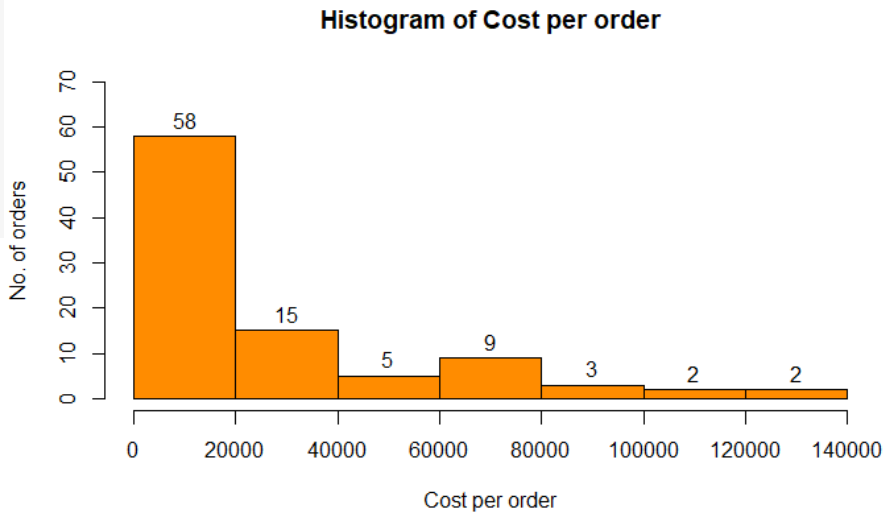
The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier, Item Description, Cost per order and Arrival Time**

ii. Generate the charts and tables to view the frequency distributions of the 4 variables of interest to the manager. (4 marks: 0.5 marks per table/chart)

Cost per Order

```
par(mar=c(5,4,4,2))
h3<-hist(PO$`Cost per order`,
         main="Histogram of Cost per order",
         xlab="Cost per order",
         ylab="No. of orders",
         col=c("darkorange"),
         xlim=c(0,140000),
         ylim=c(0,70),
         labels=TRUE)
```



QUESTION 2A: PURCHASE ORDER DASHBOARD

The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier, Item Description, Cost per order and Arrival Time**

ii. Generate the charts and tables to view the frequency distributions of the 4 variables of interest to the manager. (4 marks: 0.5 marks per table/chart)

Cost per Order

```
# extract frequency table from hist()
Cost.Group<-cut(PO$`Cost per order`,h3$breaks)
t3<-table(Cost.Group)
kable(t3, caption = "Frequency distribution for Cost per order")
```

Frequency distribution for Cost per order

Cost.Group	Freq
(0,2e+04]	58
(2e+04,4e+04]	15
(4e+04,6e+04]	5
(6e+04,8e+04]	9
(8e+04,1e+05]	3
(1e+05,1.2e+05]	2
(1.2e+05,1.4e+05]	2

QUESTION 2A: PURCHASE ORDER DASHBOARD

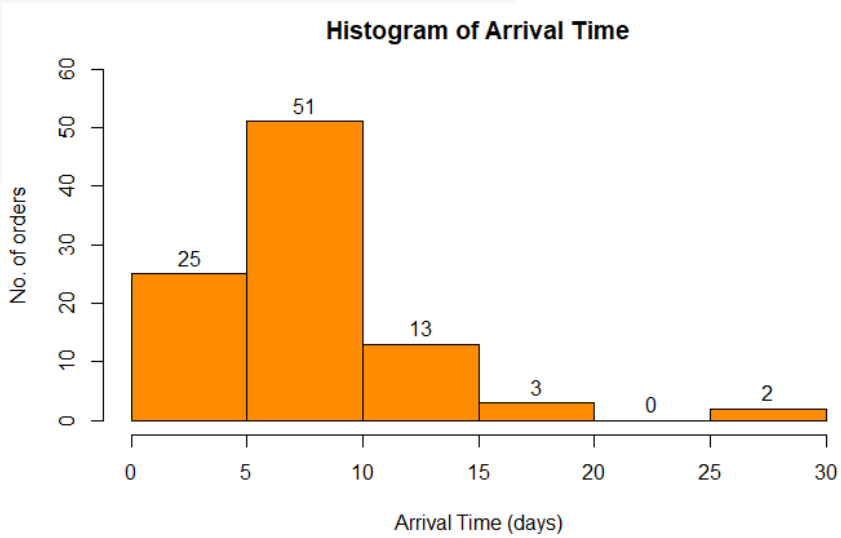
The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier, Item Description, Cost per order and Arrival Time**

ii. Generate the charts and tables to view the frequency distributions of the 4 variables of interest to the manager. (4 marks: 0.5 marks per table/chart)

Arrival Time

```
# Arrival Time
# create histogram
h4<-hist(PO$`Arrival Time`,
         main="Histogram of Arrival Time",
         xlab="Arrival Time (days)",
         ylab="No. of orders",
         col=c("darkorange"),
         ylim = c(0, 60),
         labels=TRUE)
```



Many orders arrived between 5-10 days (50 out of 94) with most taking less than 10 days.

QUESTION 2A: PURCHASE ORDER DASHBOARD

The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier, Item Description, Cost per order and Arrival Time**

ii. Generate the charts and tables to view the frequency distributions of the 4 variables of interest to the manager. (4 marks: 0.5 marks per table/chart)

Arrival Time

```
# extract frequency table from hist()
Cost.Group<-cut(PO$`Cost per order`,h3$breaks)
t3<-table(Cost.Group)
kable(t3, caption = "Frequency distribution for Cost per order")
```

Frequency distribution for Cost per order

Cost.Group	Freq
(0,2e+04]	58
(2e+04,4e+04]	15
(4e+04,6e+04]	5
(6e+04,8e+04]	9
(8e+04,1e+05]	3
(1e+05,1.2e+05]	2
(1.2e+05,1.4e+05]	2

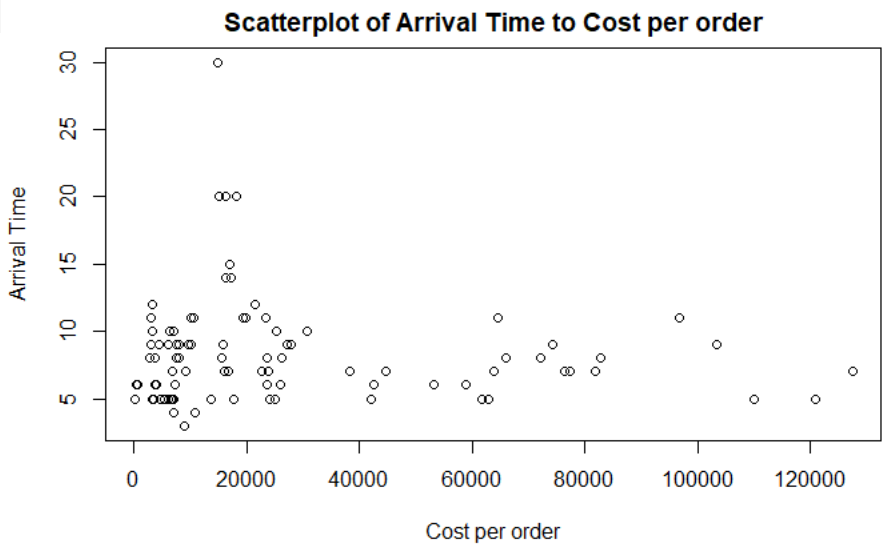
QUESTION 2A: PURCHASE ORDER DASHBOARD

The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier, Item Description, Cost per order and Arrival Time**

iii. Generate the appropriate chart to display the relationship between Cost per order and Arrival Time. (1 mark)

```
plot(PO$`Cost per order`,
     PO$`Arrival Time`,
     main="Scatterplot of Arrival Time to Cost per order",
     xlab="Cost per order",
     ylab="Arrival Time")
```



No clear observable linear relationship between Arrival Time and Cost per order.

QUESTION 2A: PURCHASE ORDER DASHBOARD

The manager would like to understand more about the items purchased in the last 3 months. More specifically, he is interested in the following purchase order information:

- **Supplier, Item Description, Cost per order and Arrival Time**

iv. Describe in your answer below your observations from each of the charts. (2.5 marks)

- Orders were most frequently from 3 Suppliers (Steelpin Inc., Hulkey Fasteners & Fast-Tie Aerospace) and least frequently from Pylon Accessories.
- Item most frequently purchased was Airframe fasteners and least frequently purchased was Panel Decal. (State most frequent and least frequent).
- Most of the orders (58 out of 94) were **in the range of $0-$20K cost per order**.
- Many orders arrived between 5-10 days (50 out of 94) with **most taking less than 10 days**.

QUESTION 2B: SUPPLIER ANALYSES DASHBOARD

The manager would like to have a deeper analyses of Supplier. In particular,

- **What are the distributions of Arrival Time for each of the Suppliers?**
- **What are the A/P terms offered by each Supplier?**

i. Display the distribution of Arrival Time in a chart for each Supplier. Differentiate each chart by including the supplier name in the chart title. The manager would like the charts to be in blue and you may use different shades of blue for each chart. (2.5 marks)

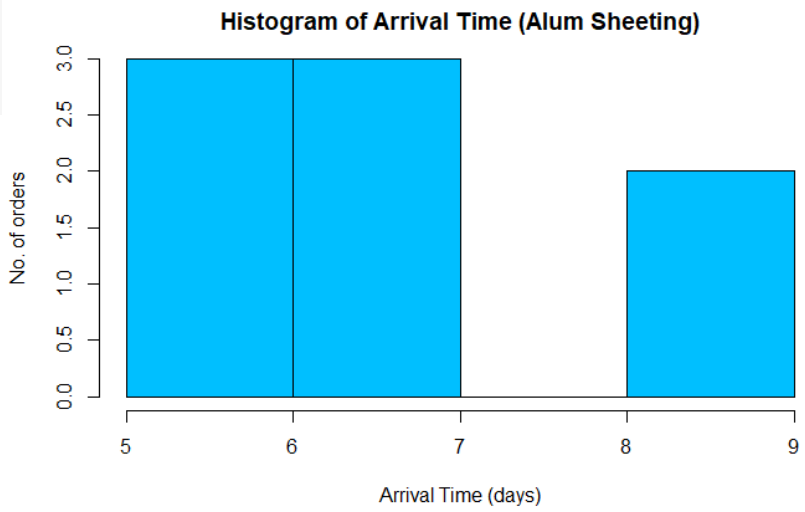
ii. From the charts, which supplier tends to ship the fastest and which tends to take the longest? Describe your answer below (1 mark)

iii. Create a table to compare the number of orders for each of the A/P Terms each supplier has. Describe your observation in your answer. (2 marks)

QUESTION 2B: SUPPLIER ANALYSES DASHBOARD

i. Display the distribution of Arrival Time in a chart for each Supplier. Differentiate each chart by including the supplier name in the chart title. The manager would like the charts to be in blue and you may use different shades of blue for each chart. (2.5 marks)

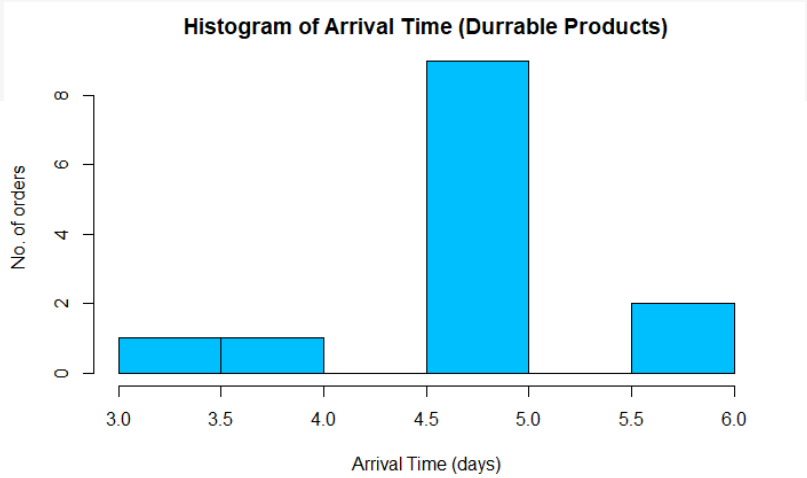
```
# create a histogram for each supplier
AS.PO<-PO %>%
  filter(Supplier=="Alum Sheeting")

hist(AS.PO$`Arrival Time`,
     main="Histogram of Arrival Time (Alum Sheeting)",
     xlab="Arrival Time (days)",
     ylab="No. of orders",
     col=c("deepskyblue"))
```



QUESTION 2B: SUPPLIER ANALYSES DASHBOARD

i. Display the distribution of Arrival Time in a chart for each Supplier. Differentiate each chart by including the supplier name in the chart title. The manager would like the charts to be in blue and you may use different shades of blue for each chart. (2.5 marks)

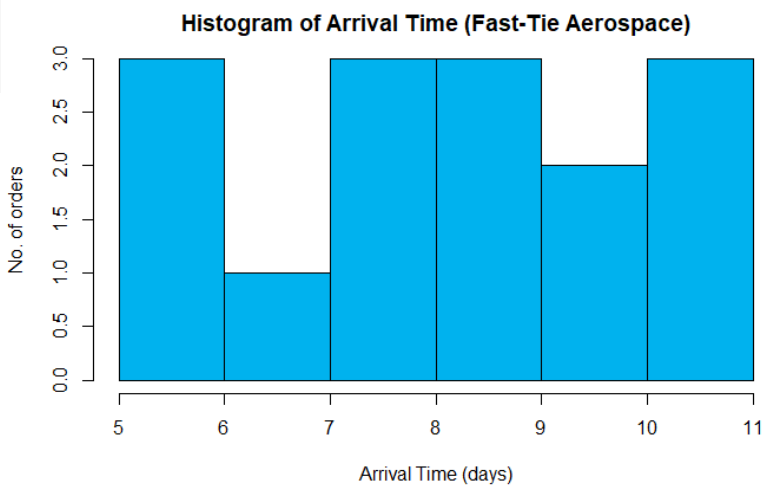
```
DP.PO<-PO %>%
  filter(Supplier=="Durrable Products")

hist(DP.PO$`Arrival Time`,
     main="Histogram of Arrival Time (Durrable Products)",
     xlab="Arrival Time (days)",
     ylab="No. of orders",
     col=c("deepskyblue1"))
```



QUESTION 2B: SUPPLIER ANALYSES DASHBOARD

i. Display the distribution of Arrival Time in a chart for each Supplier. Differentiate each chart by including the supplier name in the chart title. The manager would like the charts to be in blue and you may use different shades of blue for each chart. (2.5 marks)

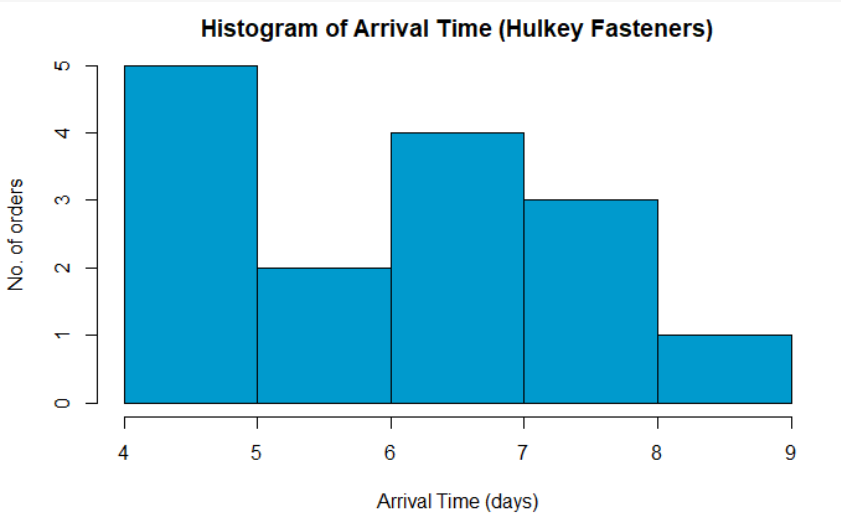
```
FT.PO<-PO %>%
  filter(Supplier=="Fast-Tie Aerospace")

hist(FT.PO$`Arrival Time`,
     main="Histogram of Arrival Time (Fast-Tie Aerospace)",
     xlab="Arrival Time (days)",
     ylab="No. of orders",
     col=c("deepskyblue2"))
```



QUESTION 2B: SUPPLIER ANALYSES DASHBOARD

i. Display the distribution of Arrival Time in a chart for each Supplier. Differentiate each chart by including the supplier name in the chart title. The manager would like the charts to be in blue and you may use different shades of blue for each chart. (2.5 marks)

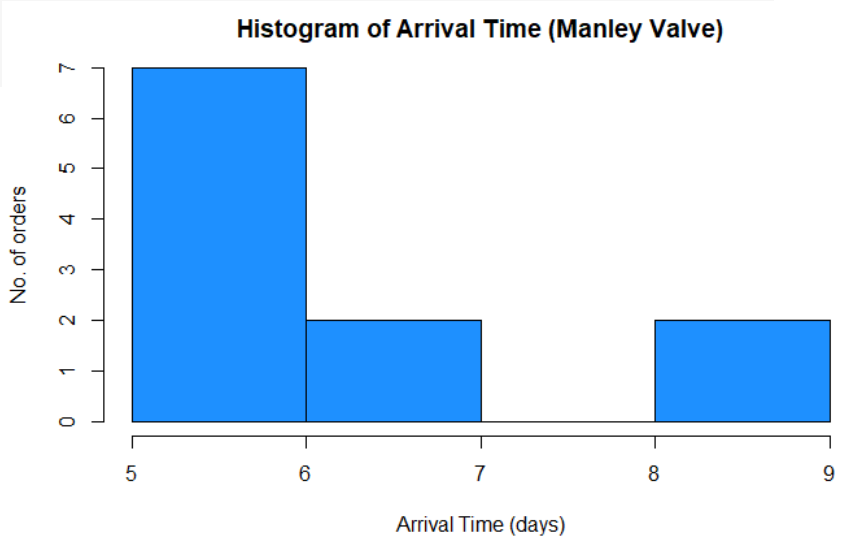
```
HF.PO<-PO %>%
  filter(Supplier=="Hulkey Fasteners")

hist(HF.PO$`Arrival Time`,
     main="Histogram of Arrival Time (Hulkey Fasteners)",
     xlab="Arrival Time (days)",
     ylab="No. of orders",
     col=c("deepskyblue3"))
```



QUESTION 2B: SUPPLIER ANALYSES DASHBOARD

i. Display the distribution of Arrival Time in a chart for each Supplier. Differentiate each chart by including the supplier name in the chart title. The manager would like the charts to be in blue and you may use different shades of blue for each chart. (2.5 marks)

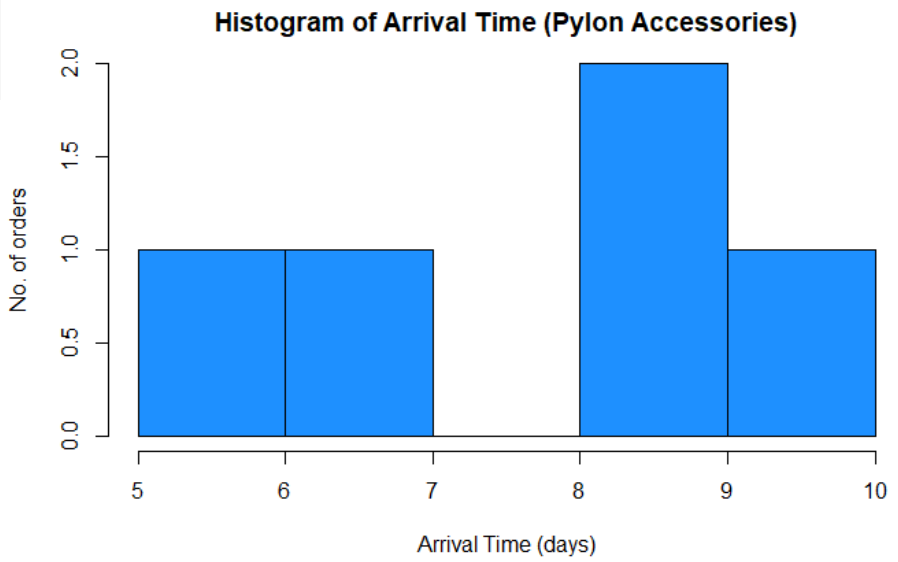
```
MV.PO<-PO %>%
  filter(Supplier=="Manley Valve")

hist(MV.PO$`Arrival Time`,
     main="Histogram of Arrival Time (Manley Valve)",
     xlab="Arrival Time (days)",
     ylab="No. of orders",
     col=c("dodgerblue"))
```



QUESTION 2B: SUPPLIER ANALYSES DASHBOARD

i. Display the distribution of Arrival Time in a chart for each Supplier. Differentiate each chart by including the supplier name in the chart title. The manager would like the charts to be in blue and you may use different shades of blue for each chart. (2.5 marks)

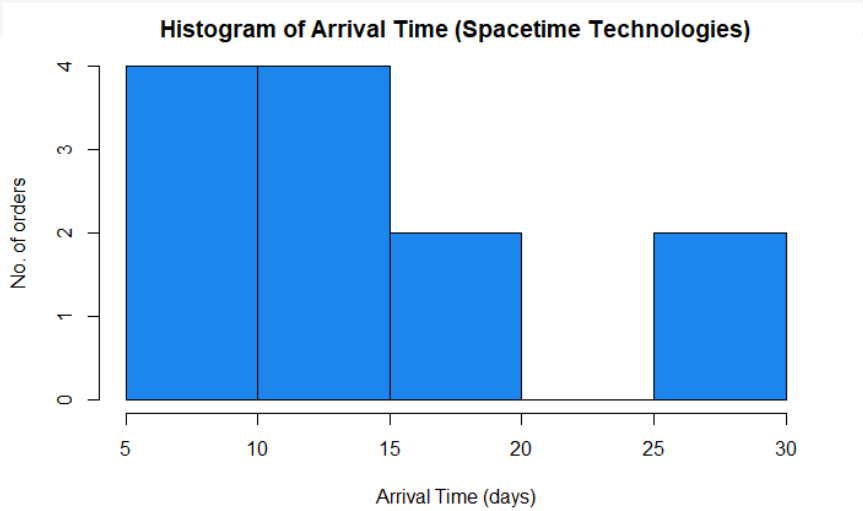
```
filter(Supplier=="Pylon Accessories")

st(PA.PO$`Arrival Time`,
   main="Histogram of Arrival Time (Pylon Accessories)",
   xlab="Arrival Time (days)",
   ylab="No. of orders",
   col=c("dodgerblue1"))
```



QUESTION 2B: SUPPLIER ANALYSES DASHBOARD

i. Display the distribution of Arrival Time in a chart for each Supplier. Differentiate each chart by including the supplier name in the chart title. The manager would like the charts to be in blue and you may use different shades of blue for each chart. (2.5 marks)

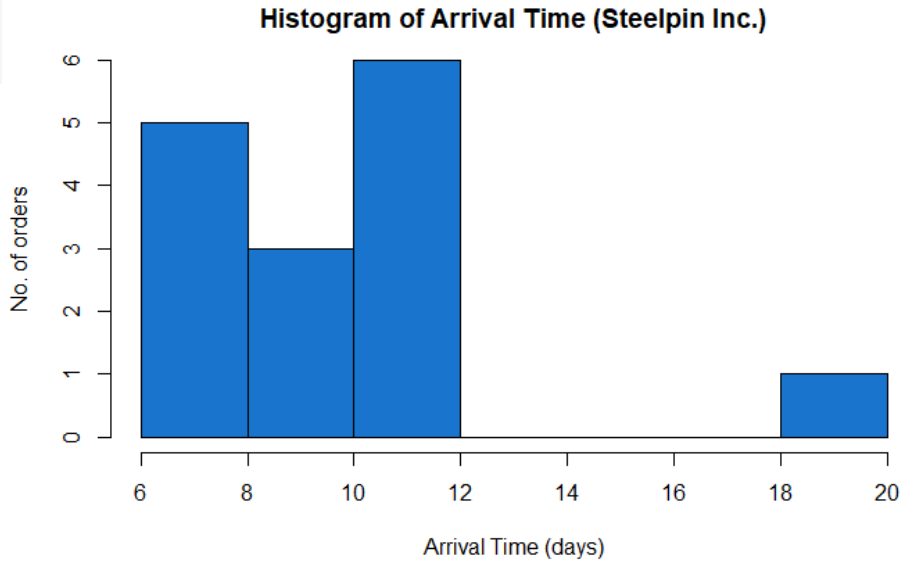
```
ST.PO<-PO %>%
  filter(Supplier=="Spacetime Technologies")

hist(ST.PO$`Arrival Time`,
     main="Histogram of Arrival Time (Spacetime Technologies)",
     xlab="Arrival Time (days)",
     ylab="No. of orders",
     col=c("dodgerblue2"))
```



QUESTION 2B: SUPPLIER ANALYSES DASHBOARD

i. Display the distribution of Arrival Time in a chart for each Supplier. Differentiate each chart by including the supplier name in the chart title. The manager would like the charts to be in blue and you may use different shades of blue for each chart. (2.5 marks)

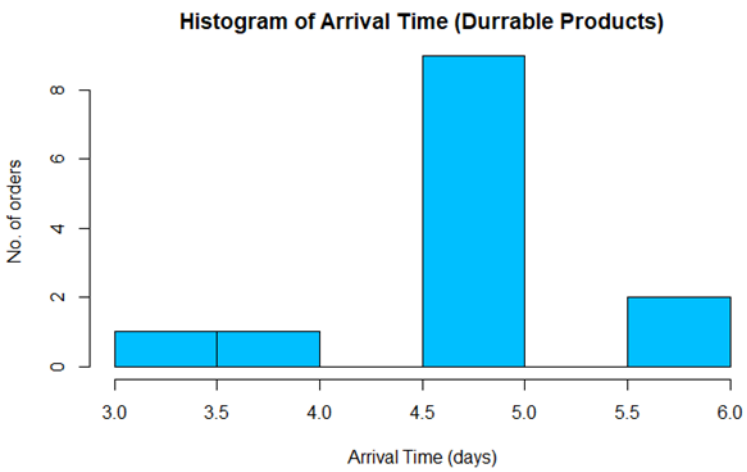
```
SI.PO<-PO %>%
  filter(Supplier=="Steelpin Inc.")

hist(SI.PO$`Arrival Time`,
     main="Histogram of Arrival Time (Steelpin Inc.)",
     xlab="Arrival Time (days)",
     ylab="No. of orders",
     col=c("dodgerblue3"))
```

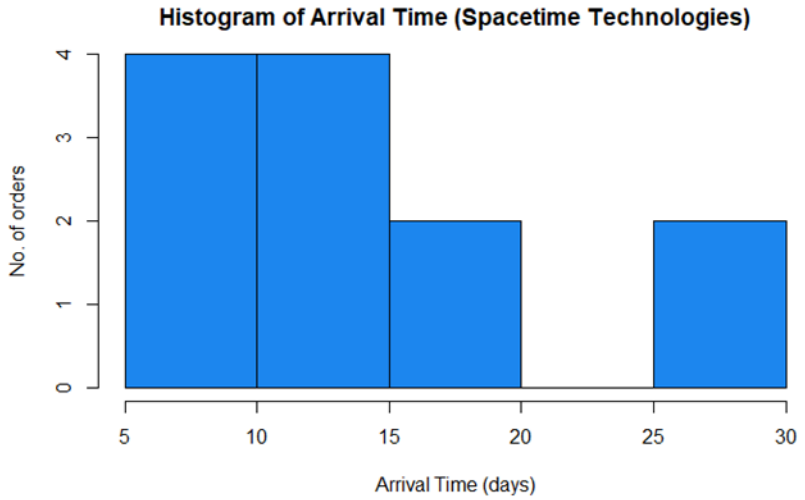


QUESTION 2B: SUPPLIER ANALYSES DASHBOARD

ii. From the charts, which supplier tends to ship the fastest and which tends to take the longest? Describe your answer below (1 mark)

- Durrable Product seems to take shortest
- Spacetime Technologies takes longest time





QUESTION 2B: SUPPLIER ANALYSES DASHBOARD

iii. Create a table to compare the number of orders for each of the A/P Terms each supplier has. Describe your observation in your answer. (2 marks)

```
PO1 <- PO %>%
  group_by(`Supplier`,`A/P Terms (Months)`) %>%
  tally()

PO1.spread<- PO1 %>%
  spread(key=`Supplier`,value=n)

PO1.spread[is.na(PO1.spread)]<-0 #convert NA to 0 value

kable(PO1.spread, caption = "Contingency Table for Supplier & A/P Terms")
```

Contingency Table for Supplier & A/P Terms

A/P Terms (Months)	Alum Sheeting	Durrable Products	Fast-Tie Aerospace	Hulkey Fasteners	Manley Valve	Pylon Accessories	Spacetime Technologies	Steelpin Inc.
15	0	0	0	0	0	5	0	0
25	0	0	0	0	0	0	12	0
30	8	0	15	15	11	0	0	15
45	0	13	0	0	0	0	0	0

- Each supplier only offers one type of AP terms
- The most common AP Terms is 30 months.

QUESTION 2C: ORDERS ACROSS MONTHS DASHBOARD

i. The manager wanted to analyze the frequency of orders from each supplier for each of the 4 months. Create a new variable Month where Month will be the month of Order Date. (1 mark)

```
PO$Month<-format(as.Date(PO$`Order Date`), "%m")
```

QUESTION 2C: ORDERS ACROSS MONTHS DASHBOARD

ii. Create the appropriate chart and table for the manager to be able to visually compare the frequency of orders from each supplier across the 4 months. (2 marks)

```
PO2 <- PO %>%
  group_by(`Supplier`,`Month`) %>%
  tally()
PO2.spread<- PO2 %>%
  spread(key="Supplier",value=n)
PO2.spread[is.na(PO2.spread)]<-0 #convert NA to 0 value
kable(PO2.spread, caption = "Contingency Table for Supplier & Order Month")
```

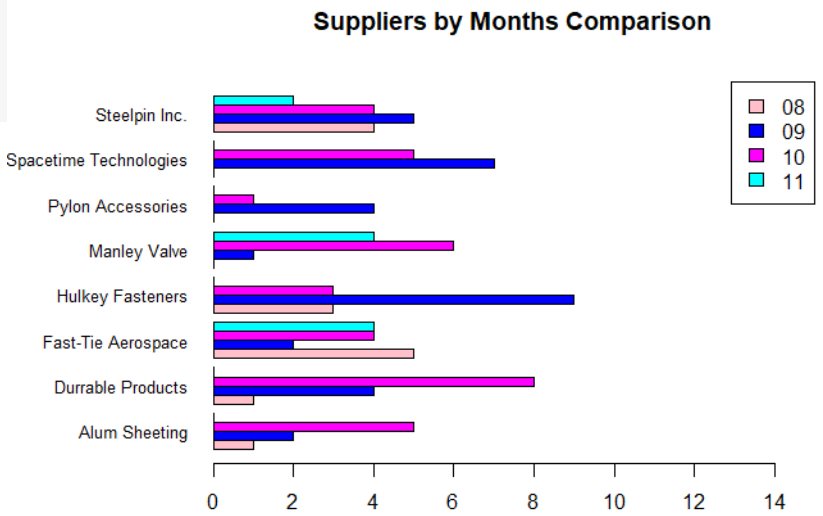
Contingency Table for Supplier & Order Month

Month	Alum Sheeting	Durrable Products	Fast-Tie Aerospace	Hulkey Fasteners	Manley Valve	Pylon Accessories	Spacetime Technologies	Steelpin Inc.
08	1	1	5	3	0	0	0	4
09	2	4	2	9	1	4	7	5
10	5	8	4	3	6	1	5	4
11	0	0	4	0	4	0	0	2

QUESTION 2C: ORDERS ACROSS MONTHS DASHBOARD

ii. Create the appropriate chart and table for the manager to be able to visually compare the frequency of orders from each supplier across the 4 months. (2 marks)

```
# plot horizontal grouped barplot
par(mar=c(2,10,5,2))
barmatrix.PO2<-as.matrix(PO2.spread[,c(2:9)])
barplot(barmatrix.PO2,
        beside = TRUE,
        horiz=TRUE,
        col =c("pink","blue", "magenta","cyan"),
        main="Suppliers by Months Comparison",
        xlim=c(0,15),
        cex.names=0.8,
        las=1)
legend("topright",
       cex=1,
       fill=c("pink","blue","magenta","cyan"),
       PO2.spread$Month)
```



QUESTION 2D: COST PER ORDER PARETO ANALYSES

i. The manager would like to conduct pareto analyses on Cost per order to understand if there is a small proportion of orders that contribute to significant amount of total cost per order. Could you help to generate the analyses? (2 marks)
ii. Describe in your answer below, the findings from your pareto analyses. (1 mark)

QUESTION 2D: COST PER ORDER PARETO ANALYSES

i. The manager would like to conduct pareto analyses on Cost per order to understand if there is a small proportion of orders that contribute to significant amount of total cost per order. Could you help to generate the analyses? (2 marks)

- Arrange cost per order in descending order
- Compute percentage of cost per order
- Computed cumulative percentage of customers from top most savings

```
#extract only the Savings column and sort in descending order
PO.cost<- PO %>%
  select (`Cost per order`)%>%
  arrange(desc(`Cost per order`))

#compute the percentage of savings over total savings
PO.cost$Percentage<-PO.cost$`Cost per order`/sum(PO.cost$`Cost per order`)

#compute cumulative percentage for Savings
PO.cost$Cumulative<-cumsum(PO.cost$Percentage)

#compute cumulative percentage of customers from top most savings
PO.cost$Cumulative.cust<-as.numeric(rownames(PO))/nrow(PO)

# compute percentage of customers with top 80% savings
37/nrow(PO)
```

[1] 0.393617

```
36/nrow(PO)
```

[1] 0.3829787

Tutorial 4

DATASET REQUIRED

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 (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 conduct some descriptive analytics on the dataset, to identify and understand any interesting patterns from the sales transaction data, and to develop dashboards to make visualization of these patterns better.

QUESTION 1A: CUSTOMER DASHBOARD

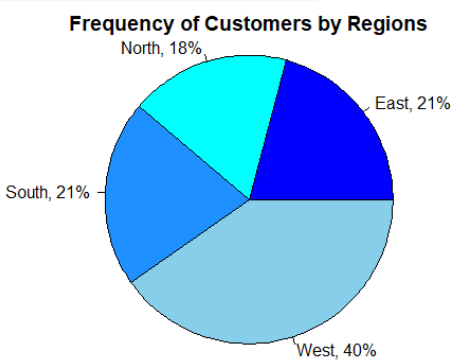
The manager would like to have a better understanding of the customer profiles. He would like the customer dashboard to be able to display the following:

i. frequency distribution for the regions the customers are from

ii. frequency distribution for the payment mode used by the customers

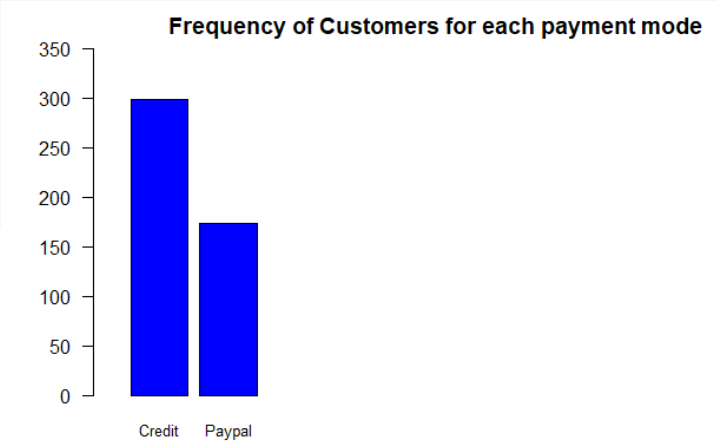
QUESTION 1A: FREQUENCY DISTRIBUTION FOR THE REGIONS THE CUSTOMERS ARE FROM

```
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 %>%
  paste(",",sep="") %>%
  paste(reg.piepercent) %>%
  paste("%",sep="")
label

pie(slice.reg,
    labels=label,
    col=c("blue","cyan","dodgerblue", "skyblue"),
    radius=1,
    main="Frequency of Customers by Regions")
```



QUESTION 1A: FREQUENCY DISTRIBUTION FOR THE PAYMENT MODE USED BY THE CUSTOMERS

```
Freq.pay<-ST %>%
count(`Payment`)
kable(Freq.pay, caption = "Frequency of Customers for each payment mode")

Freqbar <- Freq.pay$n

barplot(Freqbar,
        names.arg=Freq.pay$Payment,
        col="blue",
        beside = TRUE,
        main="Frequency of Customers for each payment mode",
        cex.names = 0.8,
        xlim=c(0,11),
        ylim = c(0,350),
        horiz=F,
        las = 1)
```



QUESTION 1B: 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 statistis for Amount in a table. The manager would like to include only these statistics: n (or number of observations), mean, sd, median, skew, kurtosis.

ii. From the results in (i), do you think Amount is normally distributed? Plot the histogram and conduct the appropriate goodness of fit test to confirm.

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?

QUESTION 1B

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.

```
# 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

- describe() is part of the psych package.

QUESTION 1B

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.

```
tab.1b<- describe(ST$Amount) %>%
  as.data.frame()

tab.1b %>%
  select(-c(range, trimmed, mad, se, min, max)) %>% # remove the columns you don't want
  mutate(vars = 'Amount') %>%
  mutate(across(where(is.double), round, 2)) %>% #round the numbers to 2 decimal places
  kable(row.names = FALSE, caption = "Descriptive Statistics for `Amount`")
```

Descriptive Statistics for Amount

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

- select() is part of the dplyr package.
- mutate() is part of the dplyr package.

SELECT FUNCTION

- We used the select function in the previous slide in tandem with the filter function.
- The select () helps us to subset columns using their names & types.
- You can select (and optionally rename) variables in a data frame, using a concise mini-language that makes it easy to refer to variables based on their names.

Simple select function

```
ST %>%
  select(Product)
```

Select more than one variable

```
ST %>%
  select(Payment, Source)
```

Select a particular class

```
ST %>%
  select(where(is.numeric))
```

USEFUL SELECTION FUNCTIONS

- Select everything but
: Select range
contains() Select columns whose name contains a character string
ends_with() Select columns whose name ends with a string
everything() Select every column
matches() Select columns whose name matches a regular expression
num_range() Select columns named x1, x2, x3, x4, x5
one_of() Select columns whose names are in a group of names
starts_with() Select columns whose name starts with a character string

MUTATE FUNCTION

- mutate() adds new variables and preserves existing ones; transmute() adds new variables and drops existing ones.

- New variables overwrite existing variables of the same name. Variables can be removed by setting their value to NULL.

```
ST2 <- ST %>%
  mutate(Price=Amount+(0.20*Amount))
```

	Cust ID	Region	Payment	Transaction Code	Source	Amount	Product	Time Of Day	Price
1	10001	East	Paypal	93816545	Web	20.19	DVD	1899-12-31 22:19:00	24.228
2	10002	West	Credit	74083490	Web	17.85	DVD	1899-12-31 13:27:00	21.420
3	10003	North	Credit	64942368	Web	23.98	DVD	1899-12-31 14:27:00	28.776
4	10004	West	Paypal	70560957	Email	23.51	Book	1899-12-31 15:38:00	28.212
5	10005	South	Credit	35208817	Web	15.33	Book	1899-12-31 15:21:00	18.396
6	10006	West	Paypal	20978903	Email	17.30	DVD	1899-12-31 13:11:00	20.760
7	10007	East	Credit	80103311	Web	177.72	Book	1899-12-31 21:59:00	213.264
8	10008	West	Credit	14132683	Web	21.76	Book	1899-12-31 04:04:00	26.112
9	10009	West	Paypal	40128225	Web	15.92	DVD	1899-12-31 19:35:00	19.104
10	10010	South	Paypal	49073721	Web	23.39	DVD	1899-12-31 13:26:00	28.068
11	10011	South	Paypal	57398827	Email	24.45	Book	1899-12-31 14:17:00	29.340
12	10012	East	Credit	34400661	Web	20.39	Book	1899-12-31 01:01:00	24.468
13	10013	North	Paypal	54242587	Web	19.54	DVD	1899-12-31 10:04:00	23.448
14	10014	East	Credit	62597750	Web	151.67	Book	1899-12-31 09:09:00	182.004

QUESTION 1B

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.

```
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)) %>%
  kable(row.names = FALSE, caption = "Descriptive Statistics for `Amount`")
```

- summarise() is part of the dplyr package.

SUMMARISE FUNCTION

- summarise() creates a new data frame.
- It will have one (or more) rows for each combination of grouping variables; if there are no grouping variables, the output will have a single row summarising all observations in the input.
- It will contain one column for each grouping variable and one column for each of the summary statistics that you have specified.
- summarise() and summarize() are synonyms.

```
ST %>%
  group_by(Source) %>%
  summarise(n=n())
```

Source <chr>	n <int>
Email	129
Web	343

```
ST %>%
  summarise(Source=n())
```

Source <int>
472
...

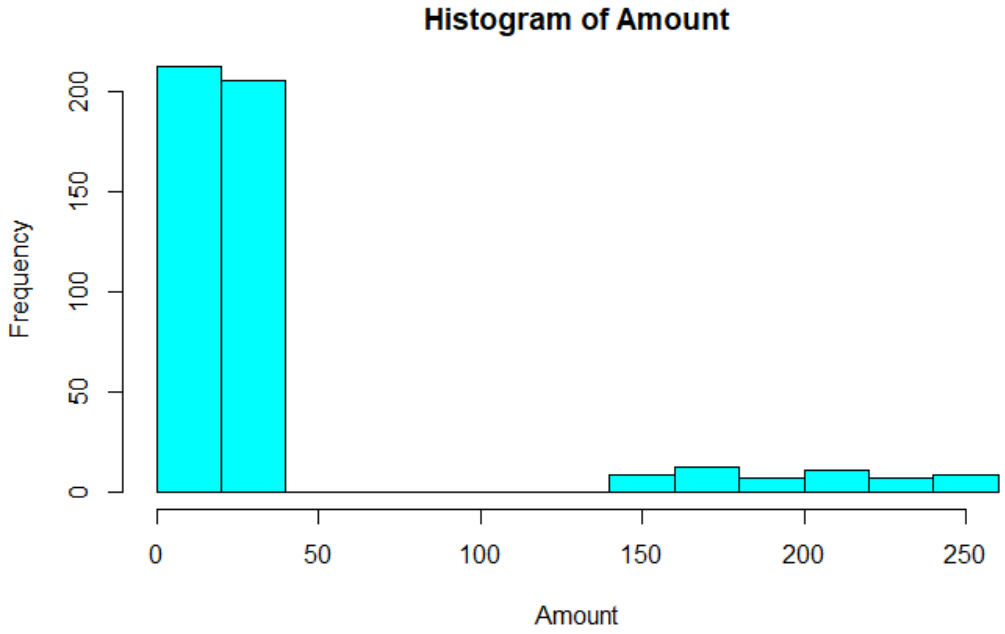
SUMMARISE FUNCTION

min(), max()	Minimum and maximum values
mean()	Mean value
median()	Median value
sum()	Sum of values
var, sd()	Variance and standard deviation of a vector
first()	First value in a vector last() Last value in a vector
nth()	Nth value in a vector
n()	The number of values in a vector
n_distinct()	The number of distinct values in a vector

QUESTION 1B

ii. From the results in (i), do you think Amount is normally distributed? Plot the histogram and conduct the appropriate goodness of fit test to confirm.

```
hist(ST$Amount,
     col = "cyan",
    xlab = "Amount",
    main="Histogram of Amount")
```



QUESTION 1B

ii. From the results in (i), do you think Amount is normally distributed? Plot the histogram and conduct the appropriate goodness of fit test to confirm.

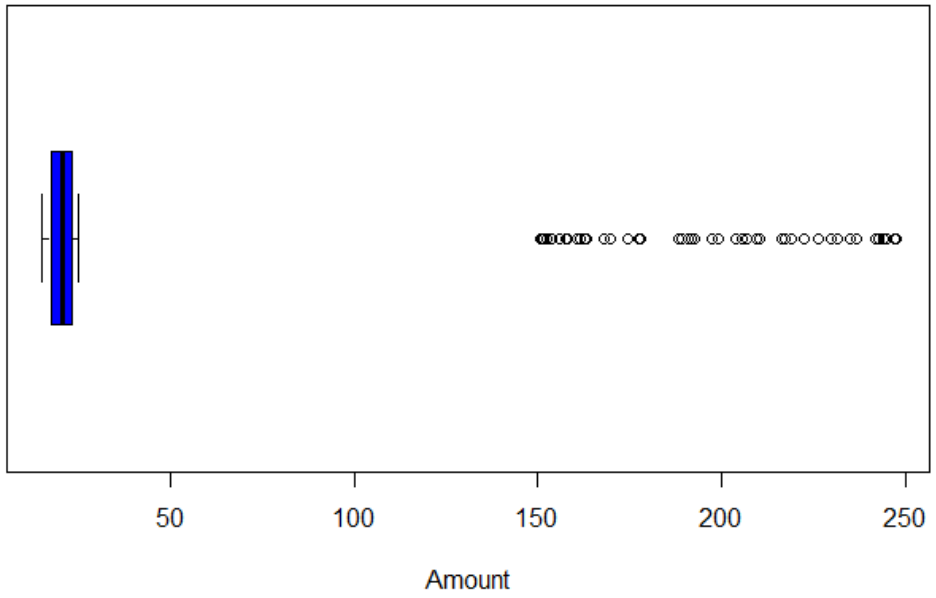
```
# Shapiro Wilkin Test
shapiro.test(ST$Amount)
```

```
        Shapiro-Wilk normality test

data:  ST$Amount
W = 0.42617, p-value < 2.2e-16
```

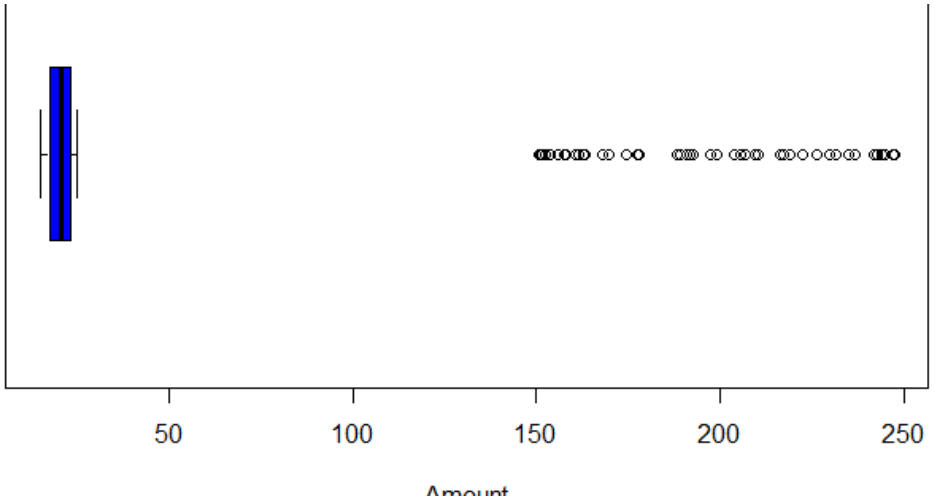
QUESTION 1B

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

```
boxplot(ST$Amount,
        range=3,
        horizontal=TRUE,
        col = "blue",
        xlab = "Amount")
```




```
boxplot(ST$Amount,
        range=1.5,
        horizontal=TRUE,
        col = "blue",
        xlab = "Amount")
```



QUESTION 1B

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.

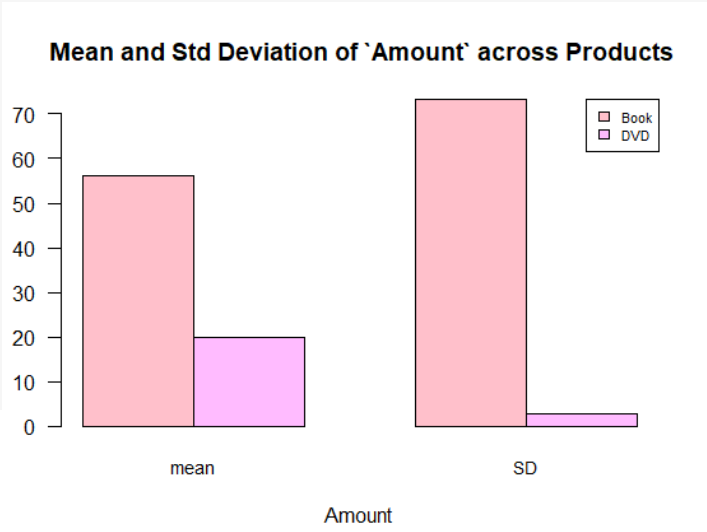
```
tab.1b2<- ST %>%
  group_by(`Product`) %>%
  summarise(mean=mean(Amount), SD=sd(Amount))
kable(tab.1b2)
```

Product	mean	SD
Book	56.21559	73.15149
DVD	19.82062	2.81961

QUESTION 1B

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.

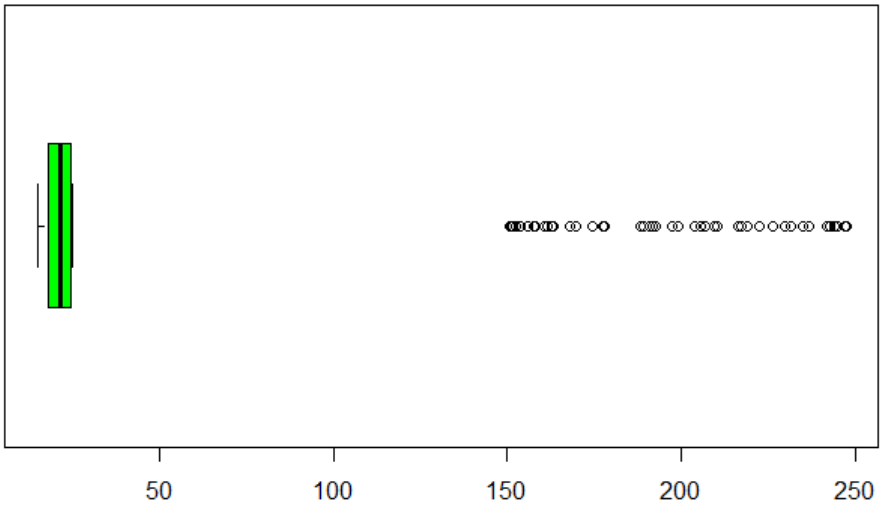
```
par(mar=c(5,10,4,2))
bar.1b2<-as.matrix(tab.1b2[,c(2:3)])
col.1b2<-c("pink","plum1")
barplot(bar.1b2,
        beside= TRUE,
        col =col.1b2,
        main=" Mean and Std Deviation of `Amount` across Products",
        cex.names=0.9,
        las=1,
        xlab="Amount")
legend("topright",
       cex=0.7,
       fill=col.1b2,
       tab.1b2$Product)
```



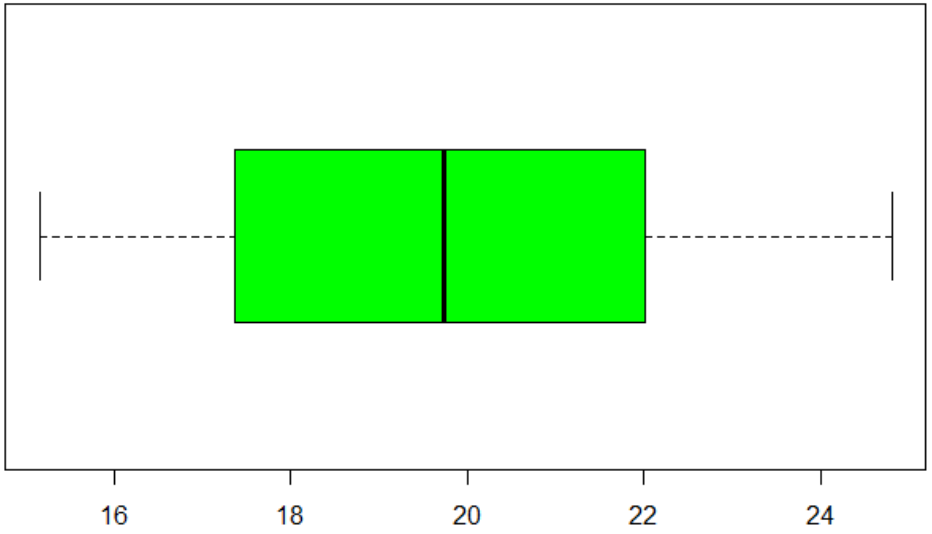
QUESTION 1B

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

```
tab.books<-ST%>%
  filter(Product=="Book")
boxplot(tab.books$Amount,
        horizontal=TRUE,
        col = "green",
        range=3)
```




```
tab.DVD<-ST%>%
  filter(Product=="DVD")
boxplot(tab.DVD$Amount,
        horizontal=TRUE,
        col = "green",
        range=3)
```



QUESTION 1B

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

- In this question, we will probably not remove the outliers. from the analyses, it looks like its sales related to books that skew the distribution.
- There are quite a number of sales with higher sales amount. Therefore they are unlikely to be outliers.
- The manager may share that this is due to the sales of rare/collector item books that tend to cost more. They are not outliers.

QUESTION 1C: COMPUTING PROBABILITIES

v. The manager would like to use the existing data to compute the probability of the following events:

i. Amount for sales transaction of Book is greater than $60.

ii. The sales transaction of DVD will come from the Web. Please compute the probabilities and type your answer below.

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

0.2030651

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

0.3411017

THANK YOU. SEE YOU NEXT WEEK.