Project 1

Achraf Cherkaoui

```
library(Rmisc)
library(IRdisplay)
library(plotly)
library(dplyr)
library(tidyverse)
library(readxl)
sales <- read_excel("sales_data_sample.xlsx")
attach(sales)
```

Question How many unique ORDERNUMBER values are in the data?

Answer :there are 307 unique ORDERNUMBER values in the data.

```
uniordn <- unique(ORDERNUMBER)
luniord <- length(uniordn)
luniord
```

```
## [1] 307
```

Question How many unique CUSTOMERNAME values are in the data?
Answer :there are 92 unique CUSTOMERNAME values in the data.

```
unicn <- unique(CUSTOMERNAME)
lunicn <- length(unicn)
lunicn
```

```
## [1] 92
```

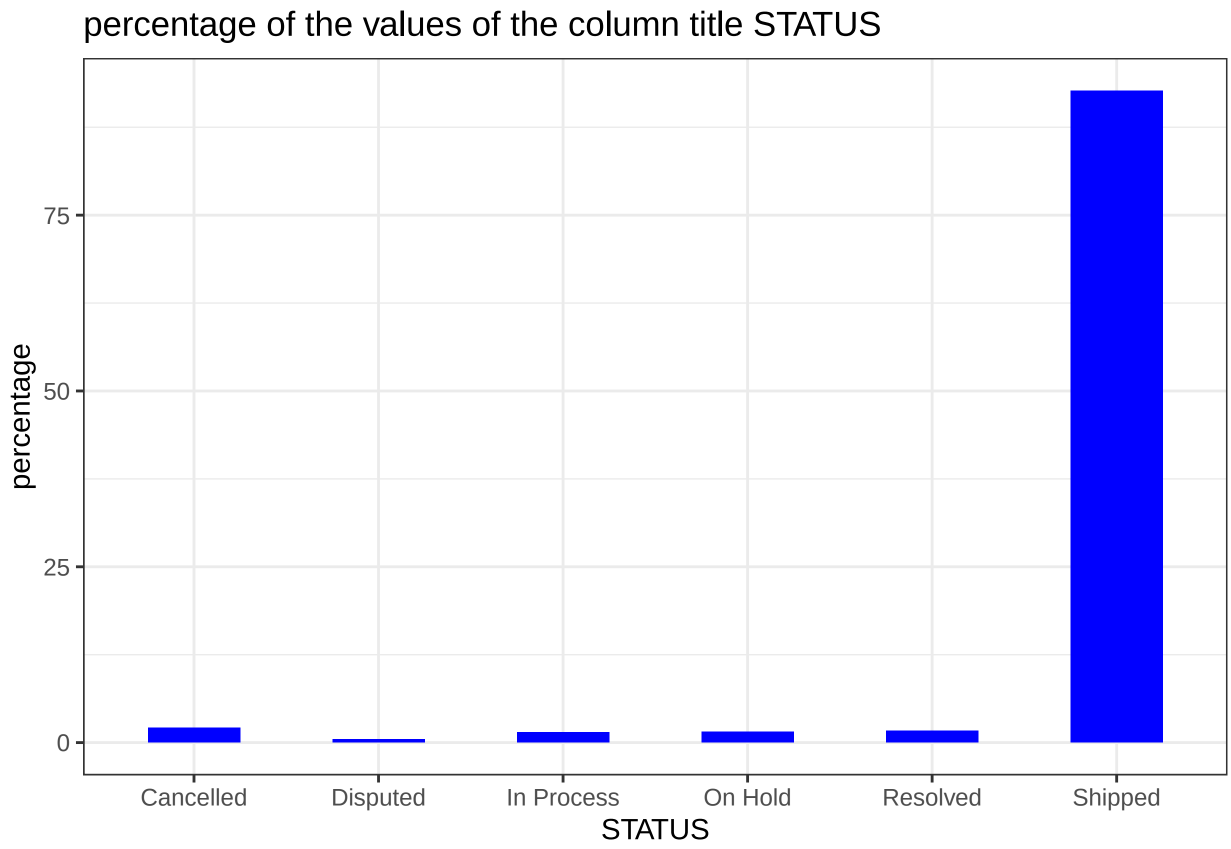
Question In a table summarize the number and percentage of the values of the column title STATUS. Visualize this information via a Pie-Chart and also via a bar chart.

Answer:

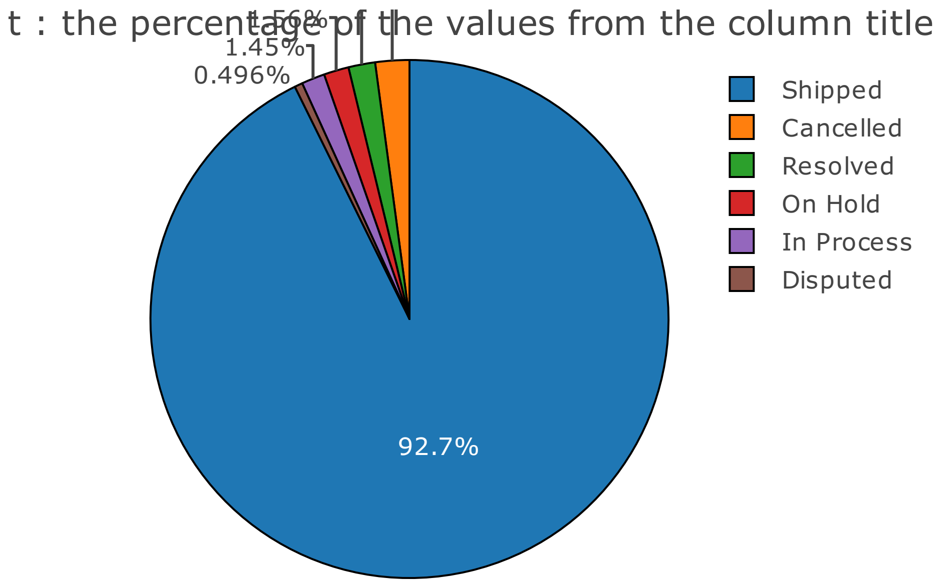
```
TabStatus <- sales %>%
  select(STATUS)%>%
  group_by(STATUS)%>%
  dplyr :: summarise( total = n()) %>%
  arrange(desc(total))%>%
  mutate( percentage1 = (total / sum(total))* 100 )
TabStatus
```

```
## # A tibble: 6 x 3
##   STATUS     total percentage1
##   <chr>      <int>       <dbl>
## 1 Shipped     2617      92.7
## 2 Cancelled     60       2.13
## 3 Resolved      47       1.66
## 4 On Hold       44       1.56
## 5 In Process    41       1.45
## 6 Disputed      14       0.496
```

```
barstatus <- TabStatus %>%
  ggplot(aes( STATUS , percentage1  ))+
  geom_bar(stat = "identity" , width = 0.5 , fill = "blue" )+
  theme_bw()+
  labs(x = "STATUS" ,
       y = "percentage" ,
         title = "percentage of the values of the column title STATUS" )
barstatus
```



```
pie1 <- plot_ly(data = TabStatus, labels = ~STATUS, values = ~percentage1,
             type = 'pie', sort= FALSE,
            marker= list(colors=colors, line = list(color="black", width=1))) %%
            layout(title="Pie chart : the percentage of the values from the column title STATUS ")
pie1
```



Question In a table list the top five CUSTOMERNAME who had the most number of orders “Shipped”. For each, also provide information on what percentage of their total orders have “Shipped”. Visualize this information via a bar chart.
Answer:

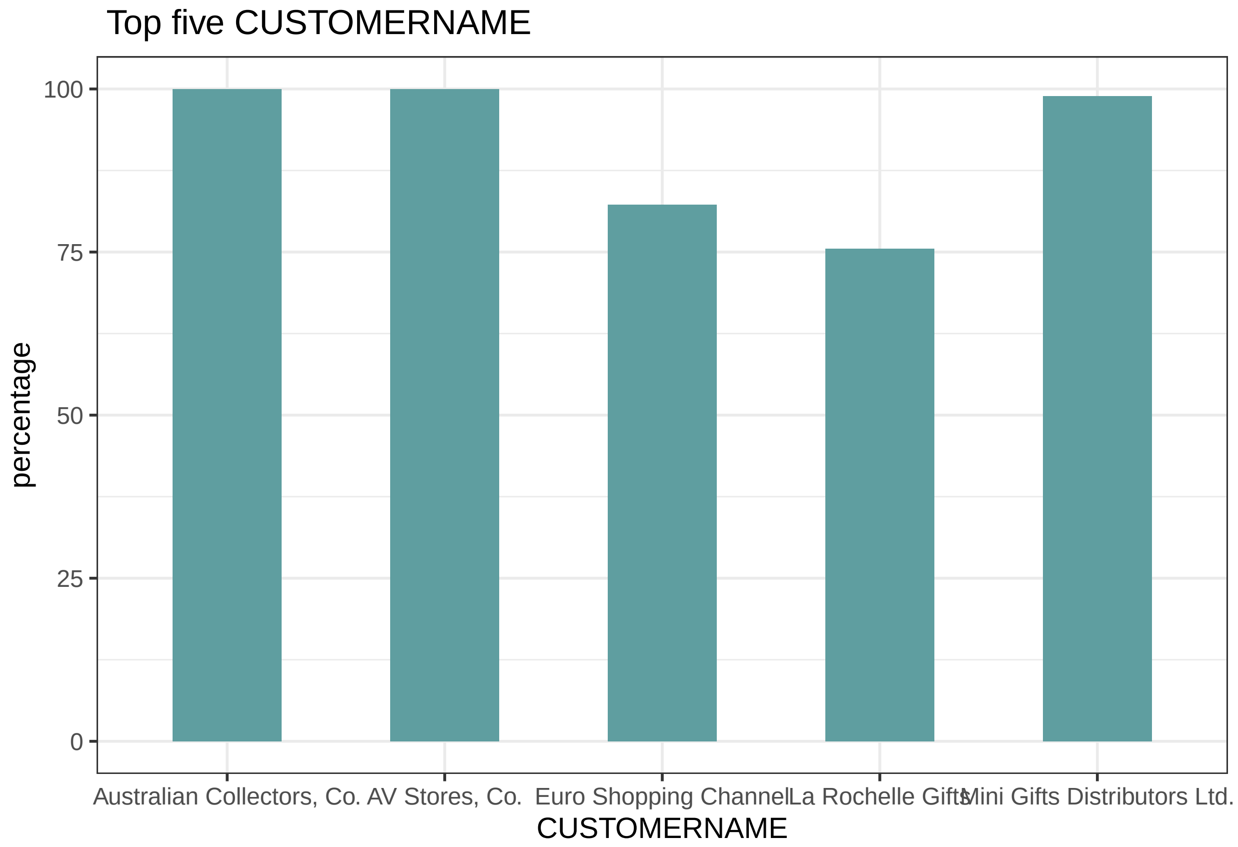
```
try1 <- sales %>%
  group_by(CUSTOMERNAME)%>%
  dplyr::summarise(totalorders= n())%>%
  arrange(desc(totalorders))

try2 <-sales%>%
  select(CUSTOMERNAME , STATUS)%>%
  filter( STATUS == "Shipped")%>%
  group_by(CUSTOMERNAME )%>%
  dplyr::summarise(shippedorders =n())%>%
  arrange(desc(shippedorders))

try3 <- try1 %>% inner_join(try2 , by="CUSTOMERNAME")%>%
  mutate( percentag2 =  (shippedorders / totalorders)*100 )%>%
  head(5)
try3
```

```
## # A tibble: 5 x 4
##   CUSTOMERNAME                 totalorders shippedorders percentag2
##   <chr>                              <int>         <int>      <dbl>
## 1 Euro Shopping Channel                259           213       82.2
## 2 Mini Gifts Distributors Ltd.         180           178       98.9
## 3 Australian Collectors, Co.              55             55       100
## 4 La Rochelle Gifts                       53             40        75.5
## 5 AV Stores, Co.                          51             51       100
```

```
bar2 <- try3 %>%
  ggplot(aes(CUSTOMERNAME , percentag2 ))+
  geom_bar(stat = "identity" , width = 0.5 , fill = "cadetblue")+
  theme_bw()+
  labs(x = "CUSTOMERNAME" ,
       y = "percentage" ,
         title = " Top five CUSTOMERNAME" )
bar2
```



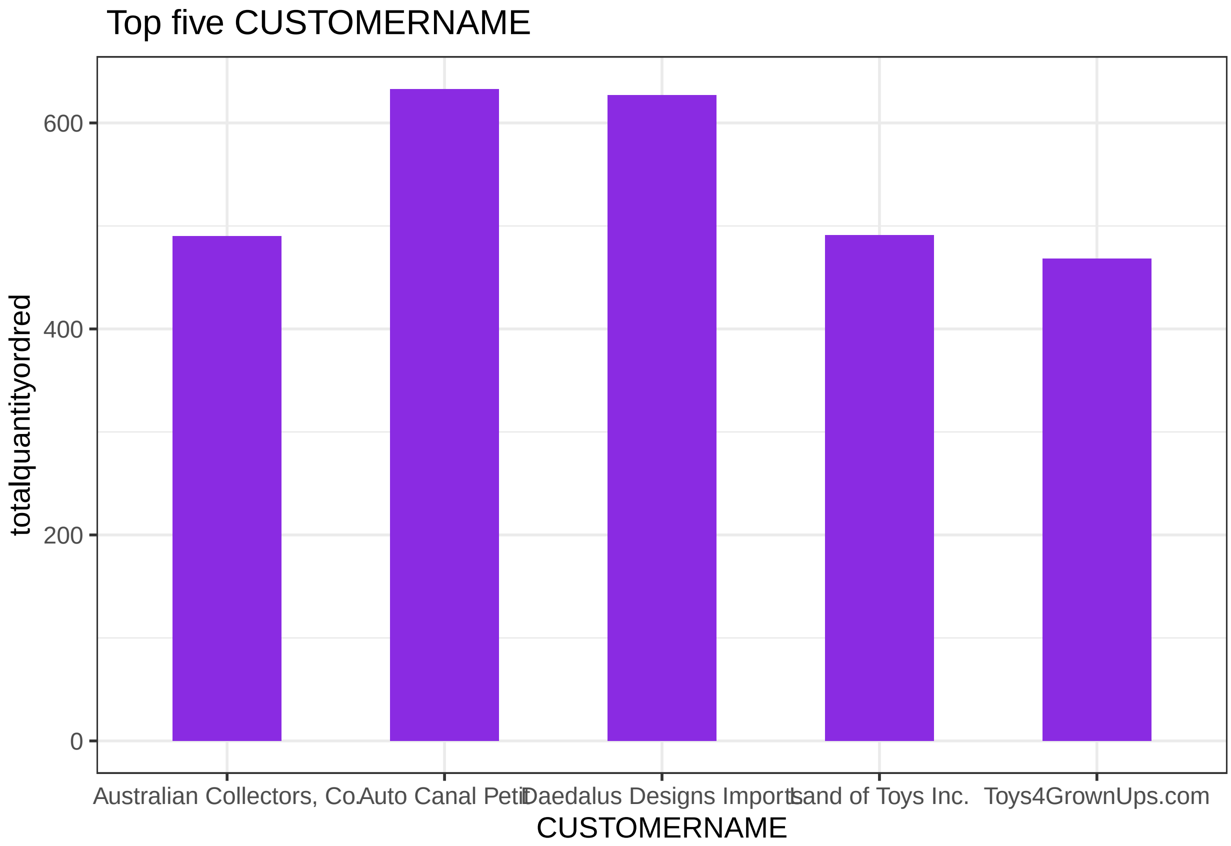
Question In a table list the top five CUSTOMERNAME who had the most number of PRODUCTLINE= 'Motorcycles' Shipped. You would need the in formation column titled QUANTITYORDERED for this. The part d above does not take into account the quantity of the motorcycles in each order. Visualize this information via a bar chart.

Answer:

```
Qe <- sales%>%
  select( CUSTOMERNAME , PRODUCTLINE , STATUS , QUANTITYORDERED)%>%
  filter( PRODUCTLINE =="Motorcycles" , STATUS == "Shipped" )%>%
  group_by(CUSTOMERNAME)%>%
  dplyr ::summarise( totalquantityordred = sum(QUANTITYORDERED))%>%
  arrange(desc(totalquantityordred))%>%
  head(5)
Qe
```

```
## # A tibble: 5 x 2
##   CUSTOMERNAME                totalquantityordred
##   <chr>                                     <dbl>
## 1 Auto Canal Petit                            633
## 2 Daedalus Designs Imports                    627
## 3 Land of Toys Inc.                           491
## 4 Australian Collectors, Co.                  490
## 5 Toys4GrownUps.com                           468
```

```
barQe <- ggplot(Qe , aes(CUSTOMERNAME , totalquantityordred))+
  geom_bar(stat = "identity", width = 0.5 , fill = "blueviolet" )+
  theme_bw()+
  labs(x = "CUSTOMERNAME" ,
       y = "totalquantityordred" ,
         title = " Top five CUSTOMERNAME" )
barQe
```



Question How many of the total 2,823 orders had STATUS value “Cancell”? Which CUSTOMERNAME had the most number of orders with STATUS value “Cancell”?

Answer:

```
TotalOrdersWithStatus <- sales %>%
  select(STATUS)%>%
  filter(STATUS == "Cancelled")

sc <- table(TotalOrdersWithStatus)
sc
```

```
## TotalOrdersWithStatus
## Cancelled
##        60
```

```
CNSVC <- sales %>%
  select( CUSTOMERNAME , STATUS)%>%
  filter( STATUS == "Cancelled")%>%
  group_by(CUSTOMERNAME)%>%
  dplyr::summarise(cancelled= n())%>%
  arrange(desc(cancelled))
CNSVC
```

```
## # A tibble: 4 x 2
##   CUSTOMERNAME            cancelled
##   <chr>                       <int>
## 1 Euro Shopping Channel          16
## 2 Scandinavian Gift Ideas        16
## 3 Land of Toys Inc.              14
## 4 UK Collectables, Ltd.          14
```

Question In a table summarize the number and percentage of the values of the column title PRODUCTLINE. Visualize this information via a Pie-Chart and also via a bar chart.

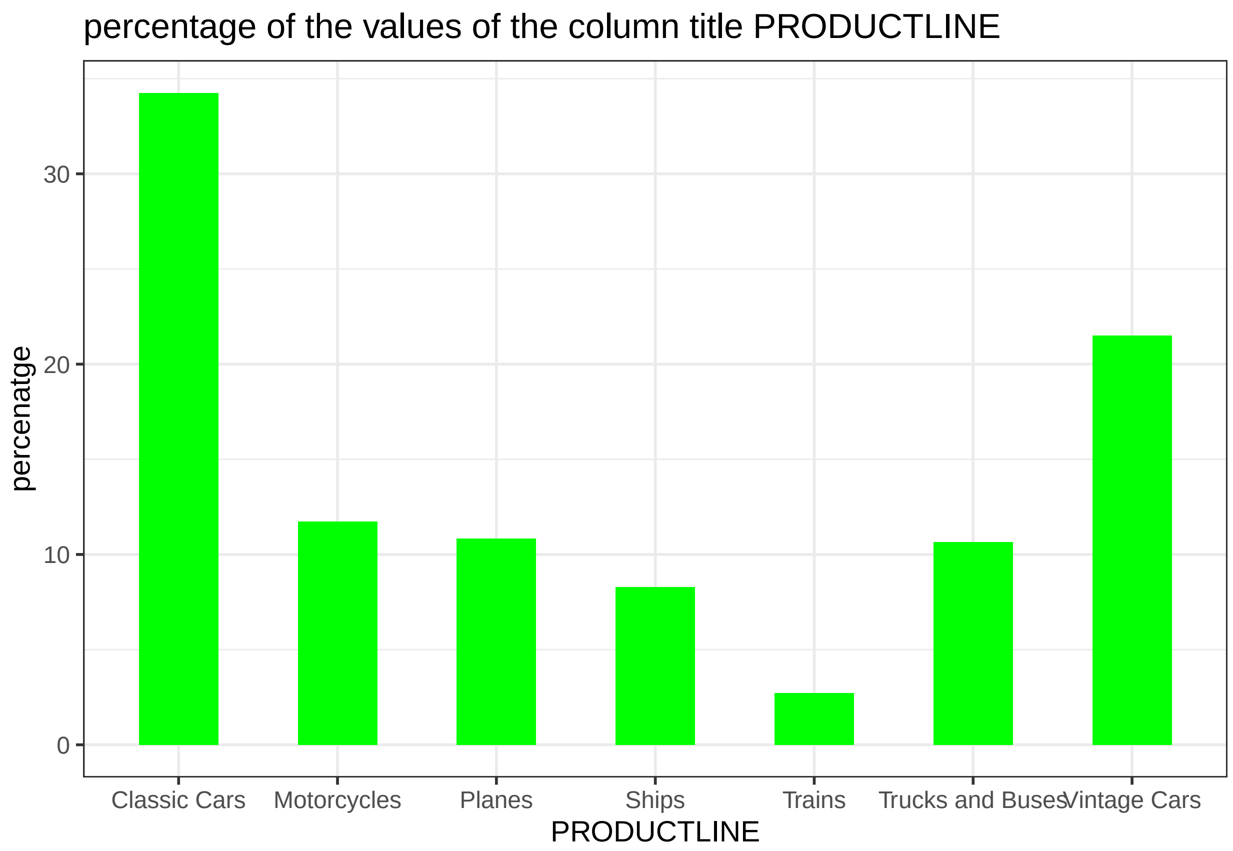
Answer:

```
productlineN <- sales%>%
  select(PRODUCTLINE)%>%
  group_by(PRODUCTLINE)%>%
  dplyr:: summarise(number = n())%>%
  mutate(percenatge = (number/sum(number))*100)
  productlineN
```

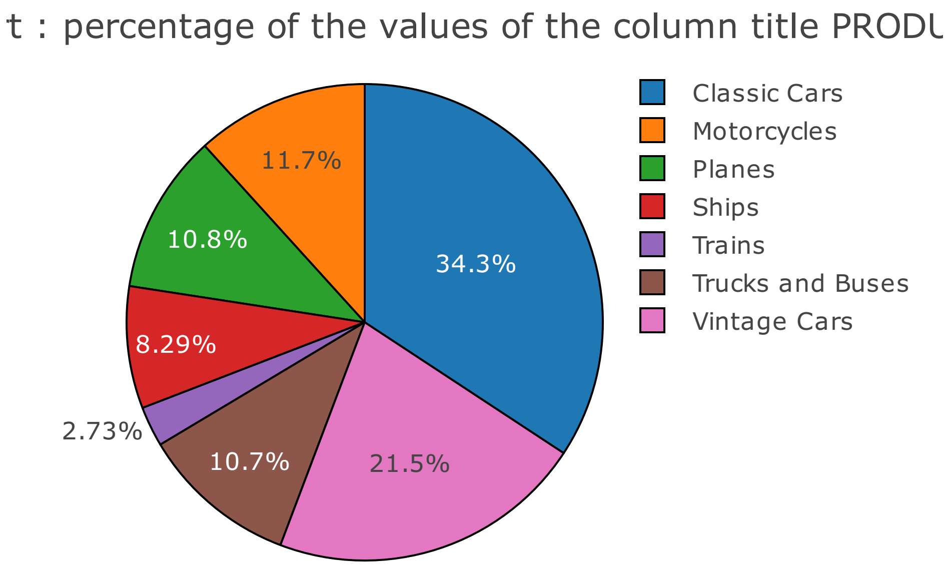
```
## # A tibble: 7 x 3
##   PRODUCTLINE      number percenatge
##   <chr>             <int>      <dbl>
## 1 Classic Cars        967      34.3
## 2 Motorcycles         331      11.7
## 3 Planes              306      10.8
## 4 Ships               234       8.29
## 5 Trains               77       2.73
## 6 Trucks and Buses    301      10.7
## 7 Vintage Cars        607      21.5
```

```
barQg <- productlineN%>%
  ggplot(aes(PRODUCTLINE , percenatge)) +
  geom_bar(stat = "identity" , width = 0.5 , fill = "green")+
   theme_bw()+
  labs(x = "PRODUCTLINE" ,
       y = "percenatge" ,
         title = "percentage of the values of the column title PRODUCTLINE " )

barQg
```



```
pie2 <- plot_ly(data = productlineN, labels = ~PRODUCTLINE, values = ~percenatge,
             type = 'pie', sort= FALSE,
            marker= list(colors=colors, line = list(color="black", width=1))) %%
            layout(title="Pie chart : percentage of the values of the column title PRODUCTLINE ")
pie2
```



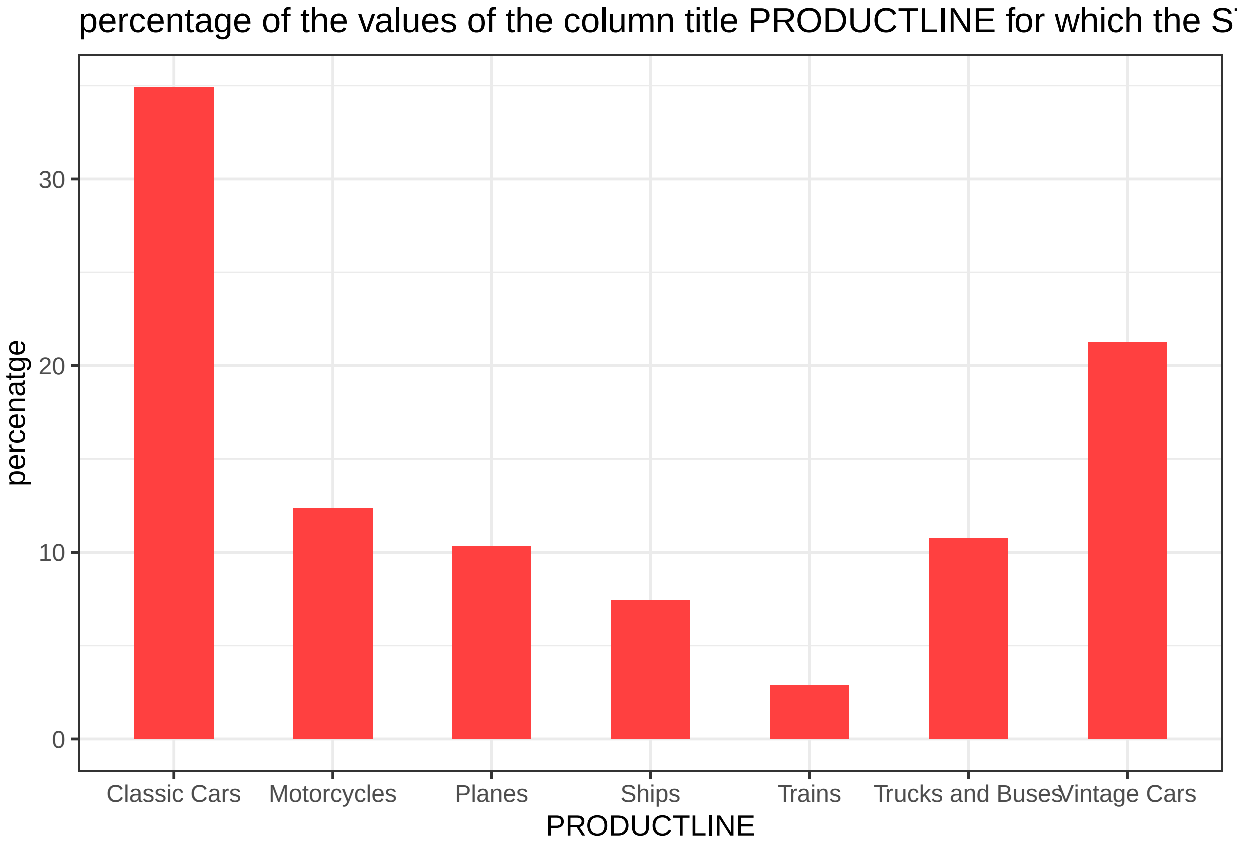
Question In a table summarize the number and percentage of the values of the column title PRODUCTLINE for which the STATUS value “Shipped”.

Answer:

```
h <- sales%>%
  select(PRODUCTLINE, STATUS)%>%
  filter(STATUS== "Shipped")%>%
  group_by(PRODUCTLINE)%>%
  dplyr::summarise(number = n())%>%
  mutate(PERCENTAGE = (number/sum(number))*100)
  h
```

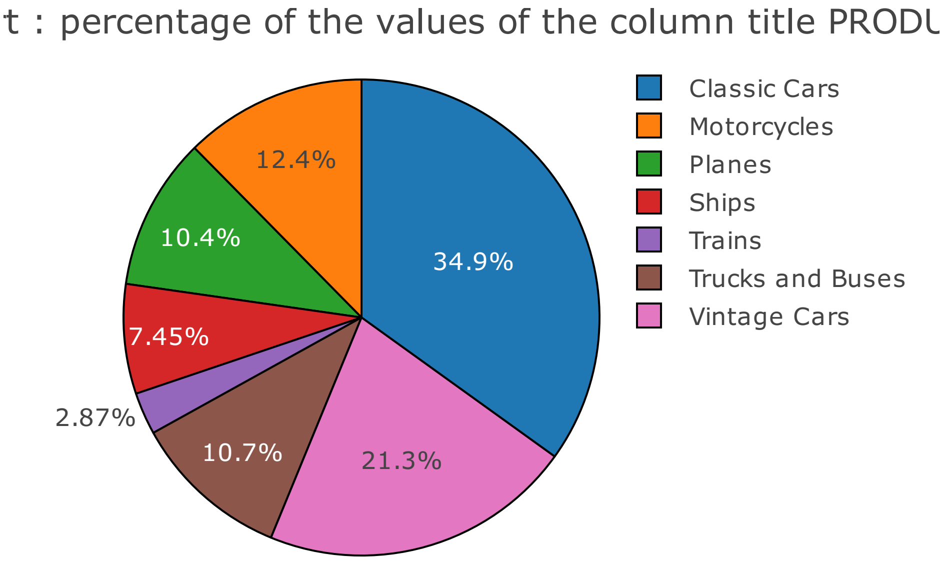
```
## # A tibble: 7 x 3
##   PRODUCTLINE      number PERCENTAGE
##   <chr>             <int>      <dbl>
## 1 Classic Cars        914      34.9
## 2 Motorcycles         324      12.4
## 3 Planes              271      10.4
## 4 Ships               195       7.45
## 5 Trains               75       2.87
## 6 Trucks and Buses    281      10.7
## 7 Vintage Cars        557      21.3
```

```
barQh <- h %>%
  ggplot(aes(PRODUCTLINE , PERCENTAGE))+
  geom_bar(stat = "identity", width = 0.5 , fill = "brown1")+
    theme_bw()+
  labs(x = "PRODUCTLINE" ,
       y = "percenatge" ,
         title = "percentage of the values of the column title PRODUCTLINE for which the STATUS value "S
barQh
```



percentage of the values of the column title PRODUCTLINE for which the S

```
pie3 <- plot_ly(data = h, labels = ~PRODUCTLINE, values = ~PERCENTAGE,
             type = 'pie', sort= FALSE,
            marker= list(colors=colors, line = list(color="black", width=1))) %%
            layout(title="Pie chart : percentage of the values of the column title PRODUCTLINE ")
pie3
```



Question What is the maximum and minimum number of motorcycles shipped in any order in USA (that is, PRODUCTLINE is Motorcycles, STATUS is Shipped and COUNTRY is USA)? Obtain a 95% confidence interval to estimate the number of motorcycles shipped in any order in USA. Explicitly state\and verify the assumptions to validate the choice of confidence interval you have chosen.

```
qq2 <- sales %>%
  select(CUSTOMERNAME , PRODUCTLINE , STATUS, COUNTRY , QUANTITYORDERED )%>%
  filter(PRODUCTLINE == "Motorcycles" , STATUS == "Shipped" , COUNTRY == "USA")%>%
   arrange(desc(QUANTITYORDERED))

summary(qq2$QUANTITYORDERED)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   26.00   34.00   33.99   42.00   50.00
```

```
CI(qq2$QUANTITYORDERED , ci=.95)
```

```
##    upper     mean    lower
## 35.47368 33.99324 32.51281
```

Answer: - the minimum is 20 motorcycles shipped in USA. - the maximum is 50 motorcycles shipped in USA. - we are 95% sure that our population mean of QUANTITYORDERED shipped motorcycles in the USA falls between 32.51 and 35.47 .