ML 1000 Assignment 2

by Anupama r.k, Queenie Tsang, Crystal (Yunan) Zhu

21/02/2021

```
##  [1] "Row.ID-0 missing values"         "Order.ID-0 missing values"
##  [3] "Order.Date-0 missing values"     "Ship.Date-0 missing values"
##  [5] "Ship.Mode-0 missing values"      "Customer.ID-0 missing values"
##  [7] "Customer.Name-0 missing values" "Segment-0 missing values"
##  [9] "Country-0 missing values"        "City-0 missing values"
## [11] "State-0 missing values"          "Postal.Code-0 missing values"
## [13] "Region-0 missing values"         "Product.ID-0 missing values"
## [15] "Category-0 missing values"       "Sub.Category-0 missing values"
## [17] "Product.Name-0 missing values"  "Sales-0 missing values"
## [19] "Quantity-0 missing values"       "Discount-0 missing values"
## [21] "Profit-0 missing values"         "diff_in_days-0 missing values"
```

Get a general idea of the data set.

```
length(unique(data$Customer.ID))
```

```
## [1] 793
```

```
#793 unique customer IDs
length(unique(data$Customer.Name))
```

```
## [1] 793
```

```
#793 unique customer names - drop one of these two vars


length(unique(data$Order.Date))
```

```
## [1] 1237
```

```
#1237 unique order dates
length(unique(data$Ship.Date))
```

```
## [1] 1334
```

```
#1334 unique ship dates - more unique ship dates than order dates - orders made on the same day were sh

length(unique(data$Segment))
```

```
## [1] 3
```

```
unique(data$Segment)
```

```
## [1] "Consumer"    "Corporate"   "Home Office"
```

```
#"Consumer"    "Corporate"   "Home Office"

unique(data$Country)
```

```
## [1] "United States"
```

```
#all are from US - could drop this variable due to no-variation introduced by it

length(unique(data$City))
```

```
## [1] 531
```

```
#531 different cities

length(unique(data$State))
```

```
## [1] 49
```

```
#49 states

length(unique(data$Postal.Code))
```

```
## [1] 631
```

```
#631 postal code - 793 unique customer IDs - some customers live very close!

unique(data$Region)
```

```
## [1] "South"   "West"    "Central" "East"
```

```
#only 4 regions

unique(data$Category)
```

```
## [1] "Furniture"       "Office Supplies" "Technology"
```

```
#only 3 categories - "Furniture" "Office Supplies" "Technology"

length(unique(data$Sub.Category))
```

```
## [1] 17
```

```
unique(data$Sub.Category)
```

```
##  [1] "Bookcases"   "Chairs"      "Labels"      "Tables"      "Storage"
##  [6] "Furnishings" "Art"         "Phones"      "Binders"     "Appliances"
## [11] "Paper"       "Accessories" "Envelopes"   "Fasteners"   "Supplies"
## [16] "Machines"    "Copiers"
```

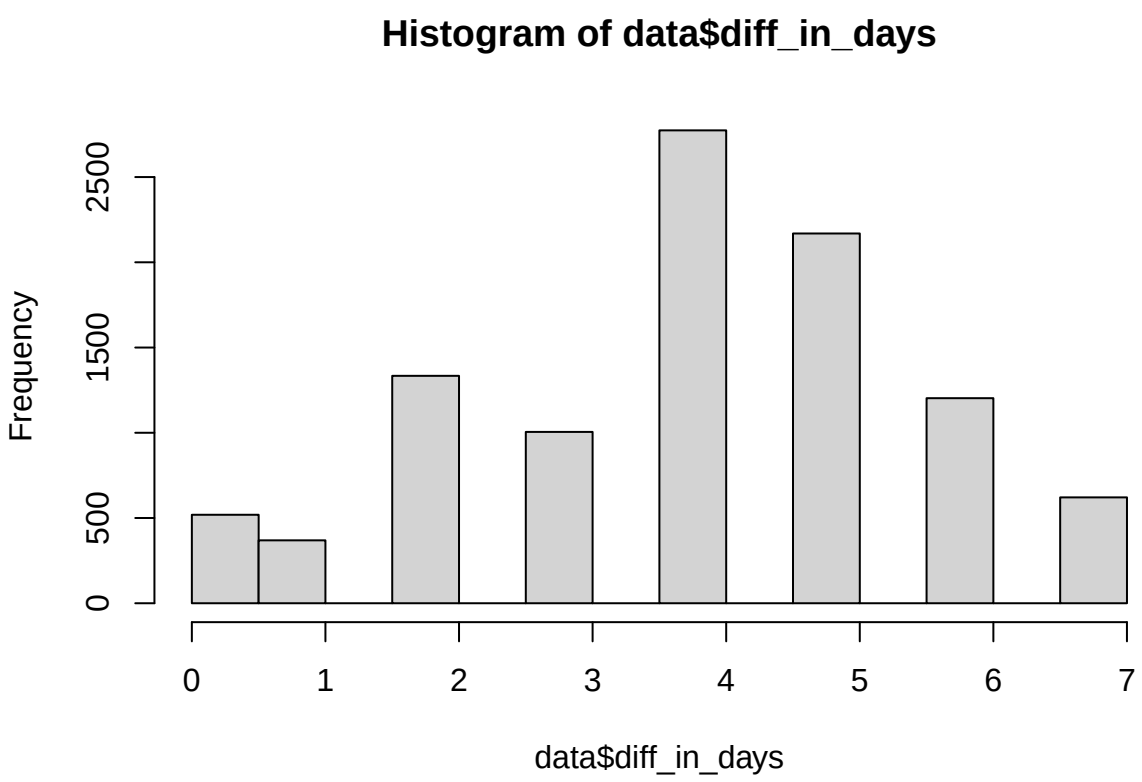
```
#17 sub-categories

length(unique(data$Product.Name))
```

```
## [1] 1850
```

```
#1850 product names
length(unique(data$Product.ID))
```

```
## [1] 1862
```

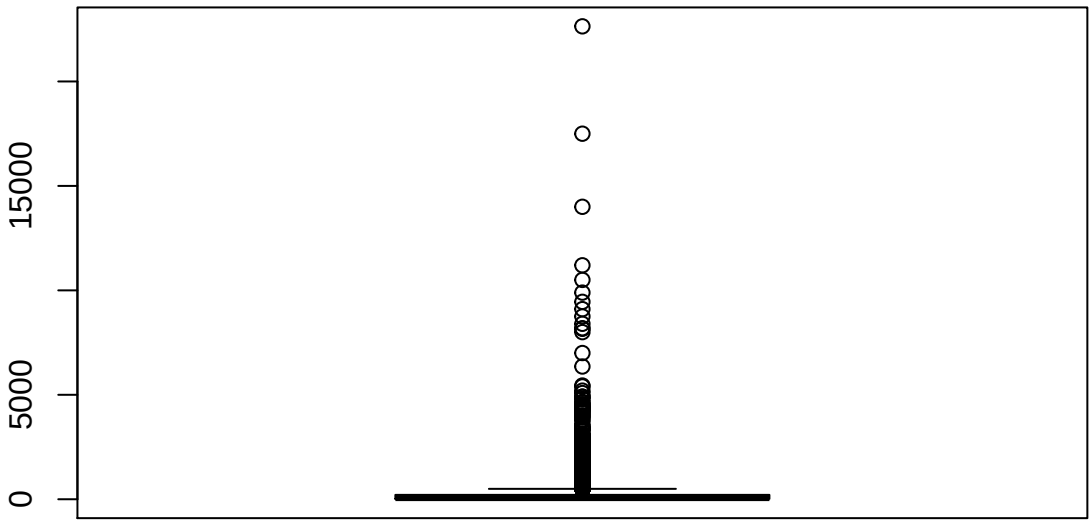
```
#1862 product IDs - potential redundant variables!

hist(data$diff_in_days)
```



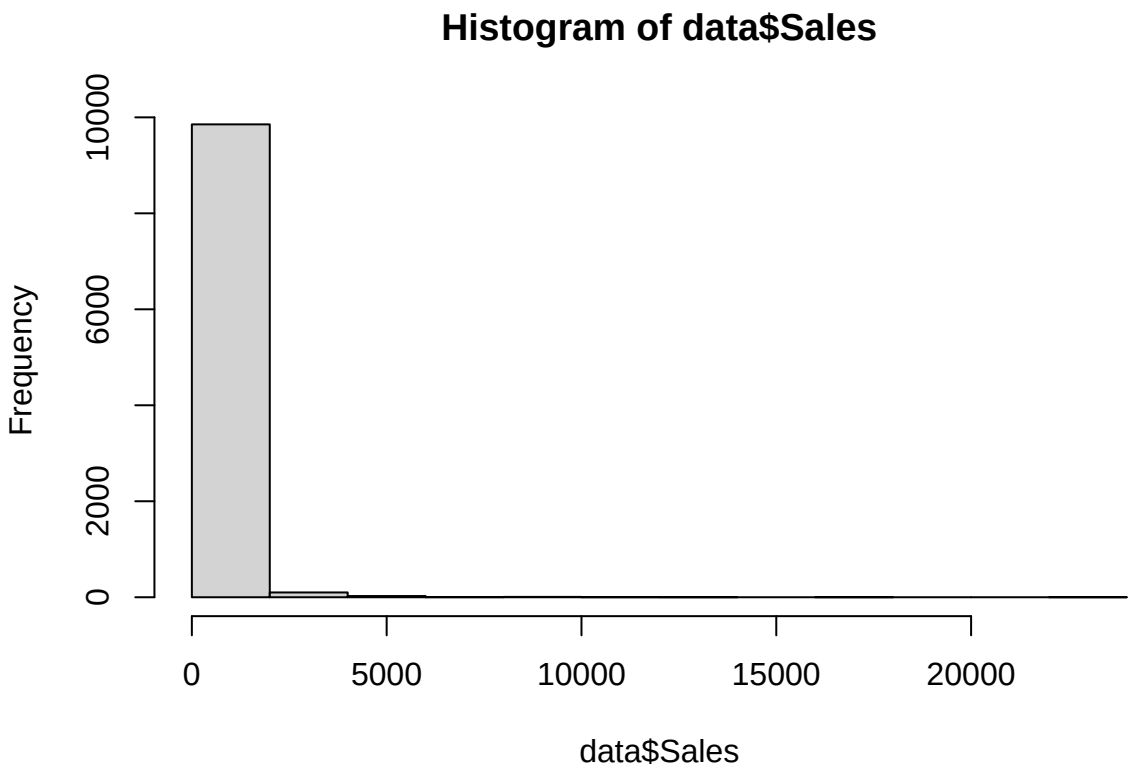

```
summary(data$Sales)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     0.444    17.280    54.490   229.858   209.940 22638.480
```

```
boxplot(data$Sales)
```



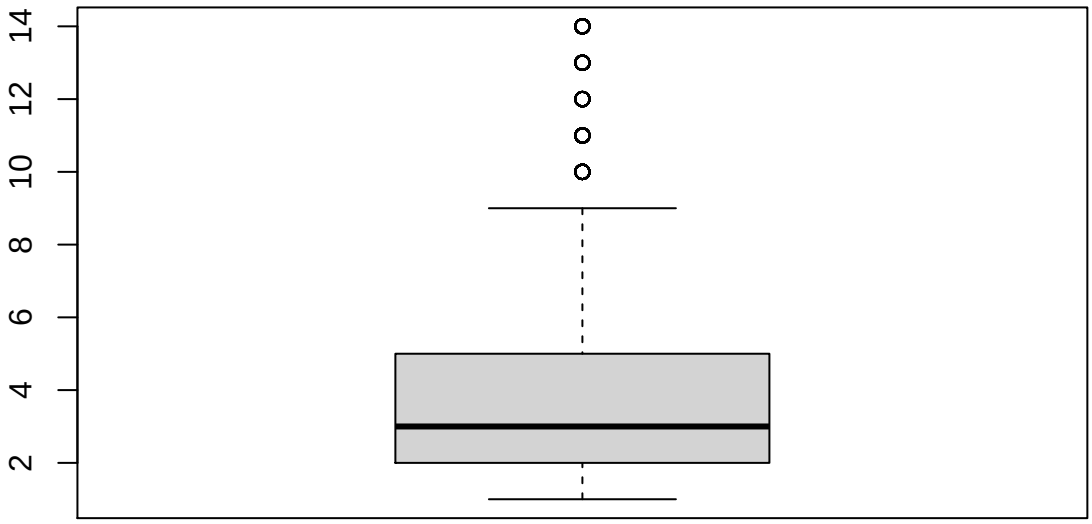

```
hist(data$Sales)
```



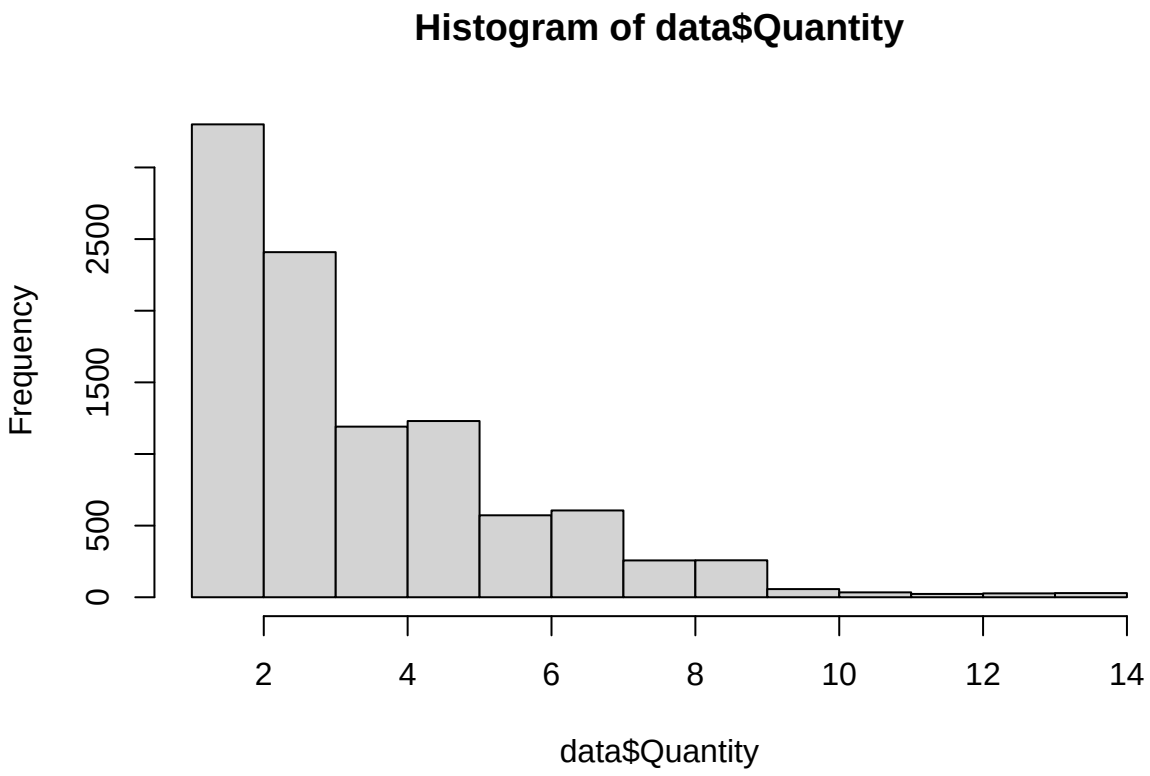

```
#a large amount of orders with very small Sales!


summary(data$Quantity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    3.00    3.79    5.00   14.00
```

```
boxplot(data$Quantity)
```



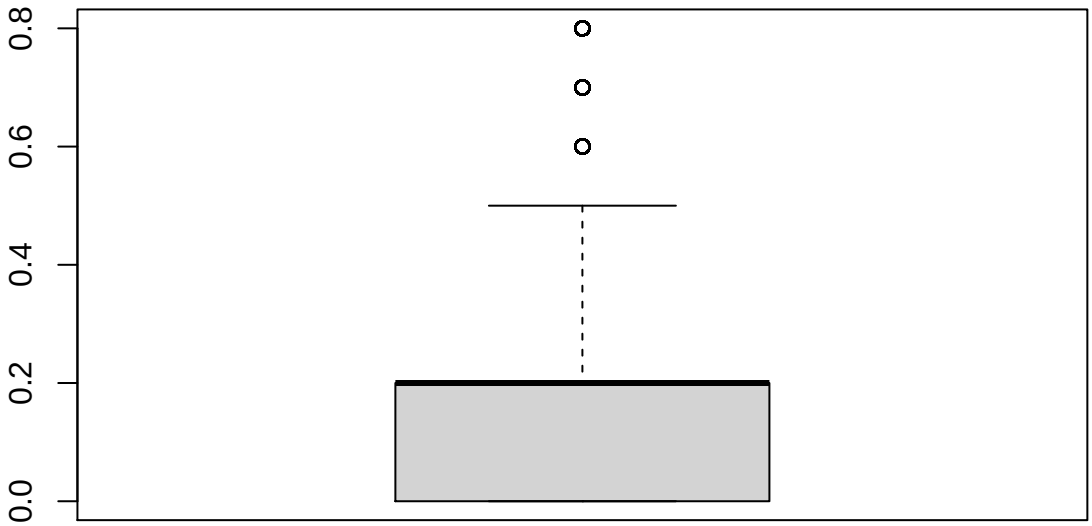

```
#not many outliers - the #of products in each order is stable?
hist(data$Quantity)
```



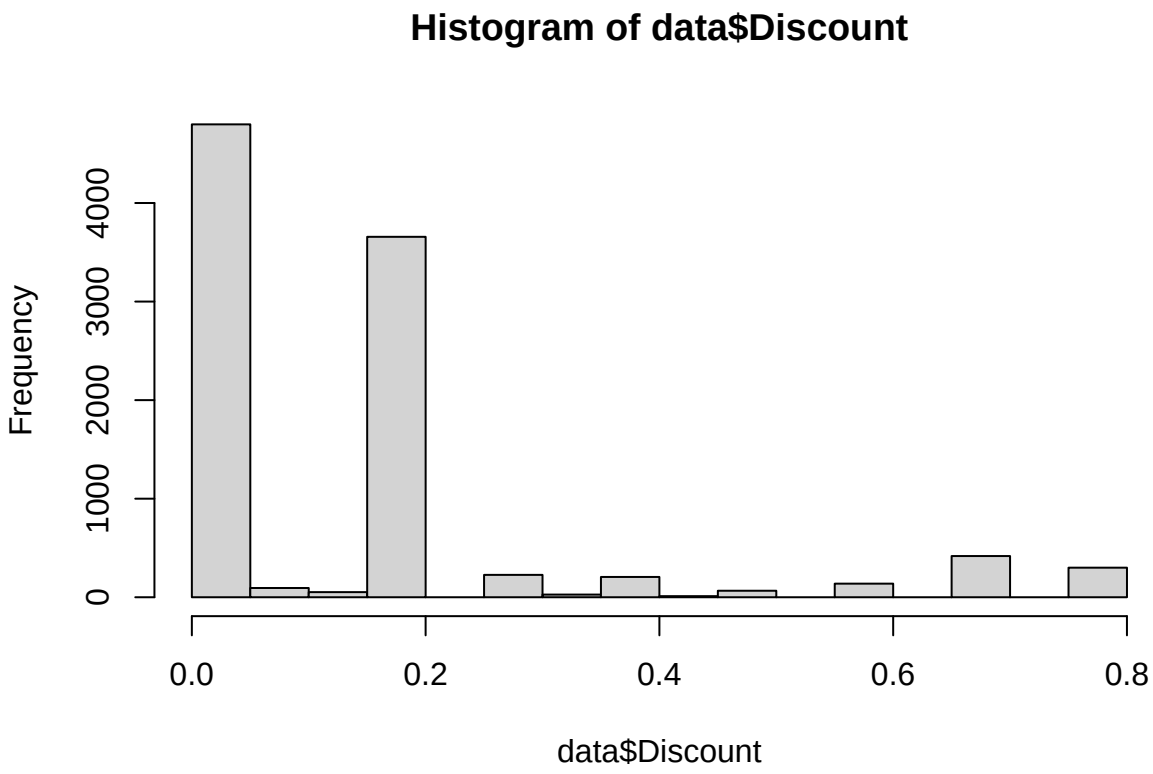

```
#very skewed distribution - most of the orders have small #of items

summary(data$Discount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.2000  0.1562  0.2000  0.8000
```

```
boxplot(data$Discount)
```



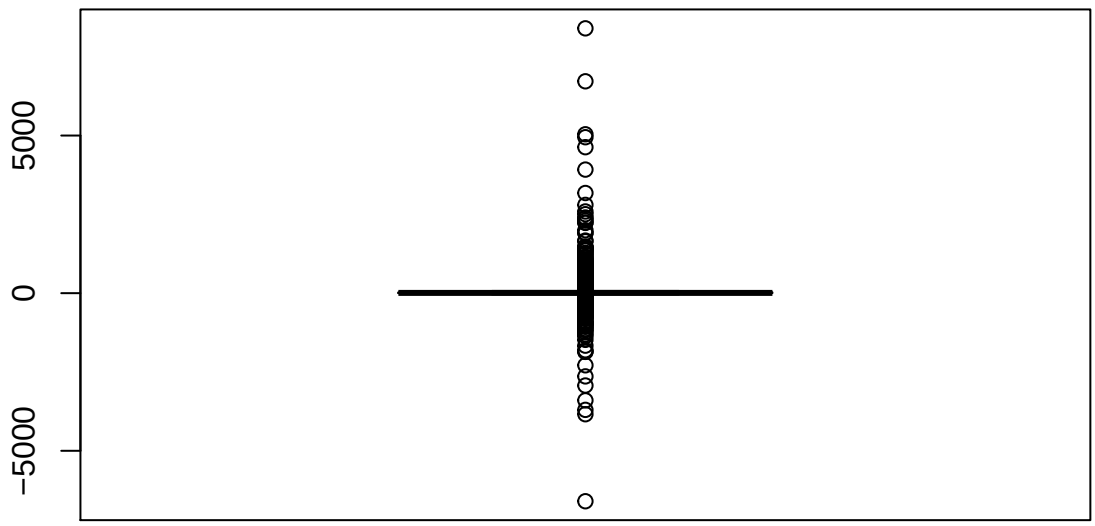

```
#a strange looking bodataplot? - median & 3rd quantile are the same (0.2) - not many orders have high d
hist(data$Discount)
```



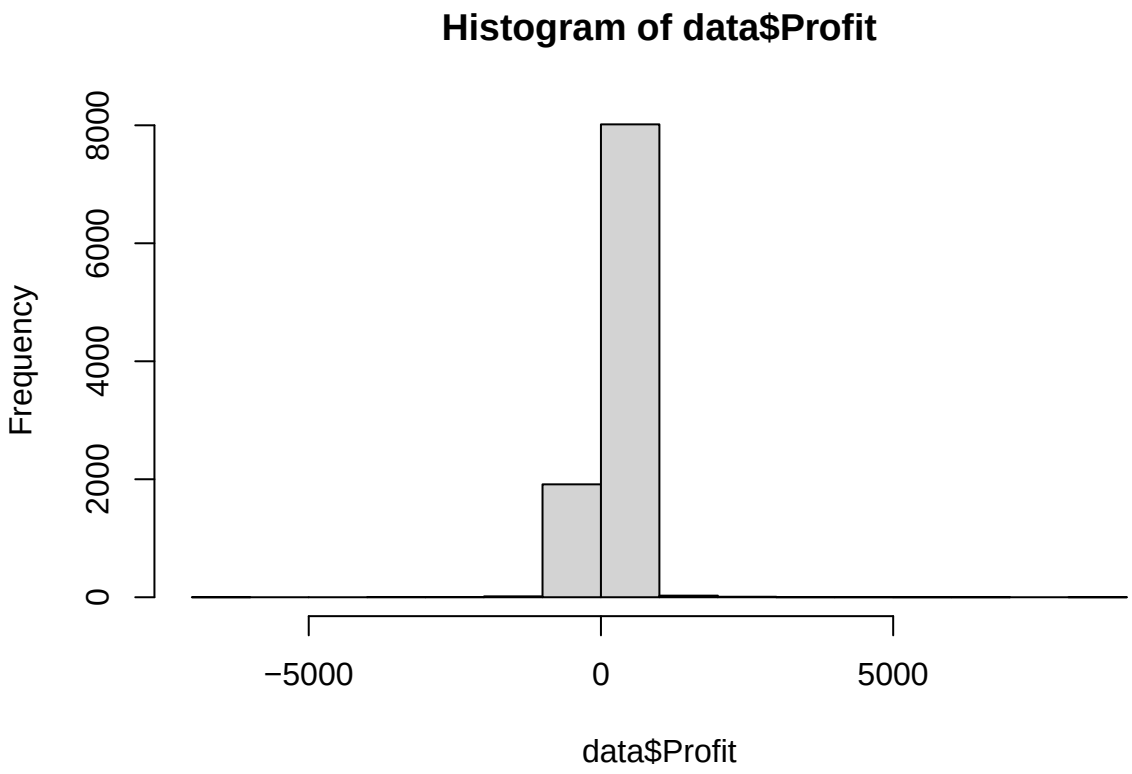

```
#most of the orders were placed without any discounts or with 20% off

summary(data$Profit)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -6599.978     1.729     8.666    28.657    29.364  8399.976
```

```
boxplot(data$Profit)
```




```
#most of the profits are outside of the box - but most of them clustered close to the box(not with so e
hist(data$Profit)
```



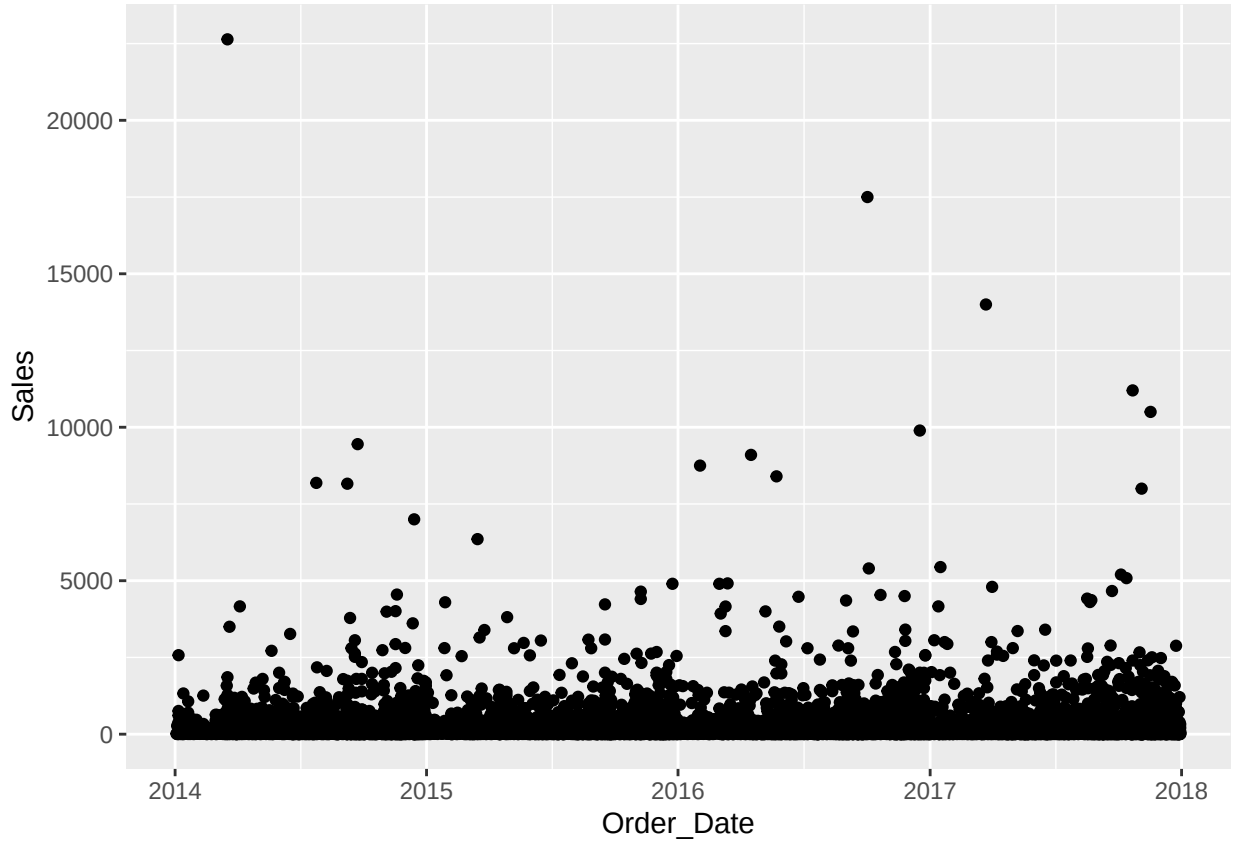

```
#most of the orders have profits ~1000 (or ~800?), and ~ -800
```

Remove the dot in the column names and replace with "_" to make variable names easier to handle:

```
##  [1] "Row_ID"        "Order_ID"      "Order_Date"    "Ship_Date"
##  [5] "Ship_Mode"     "Customer_ID"   "Customer_Name" "Segment"
##  [9] "Country"       "City"          "State"         "Postal_Code"
## [13] "Region"        "Product_ID"    "Category"      "Sub_Category"
## [17] "Product_Name"  "Sales"         "Quantity"      "Discount"
## [21] "Profit"        "diff_in_days"
```

Exploratory Data Analysis

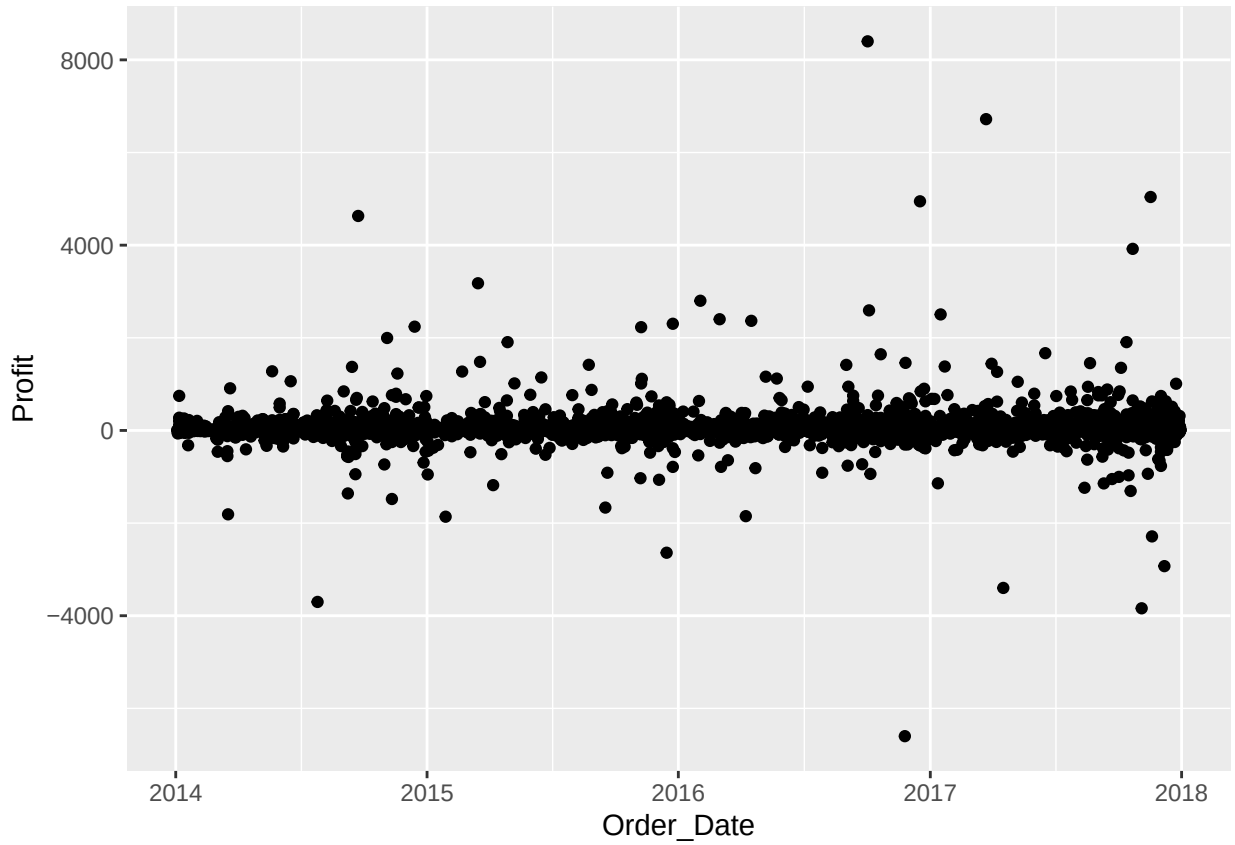
Plot Sales in relation to Order Date:



Plot Profit in relation to Order Date:

```
ggplot(data = data) +
  geom_point(mapping = aes(x = Order_Date, y = Profit), xlab="Order Date", ylab="Profit")
```

```
## Warning: Ignoring unknown parameters: xlab, ylab
```



Some outliers for certain days

```
table(data$`Sub_Category`)
```

```
##
## Accessories  Appliances         Art     Binders   Bookcases      Chairs
##         775         466         796        1523         228         617
##     Copiers   Envelopes   Fasteners Furnishings      Labels    Machines
##          68         254         217         957         364         115
##       Paper      Phones     Storage    Supplies      Tables
##        1370         889         846         190         319
```

look at the time range for these transactions, ie. start date for Order_Date column:

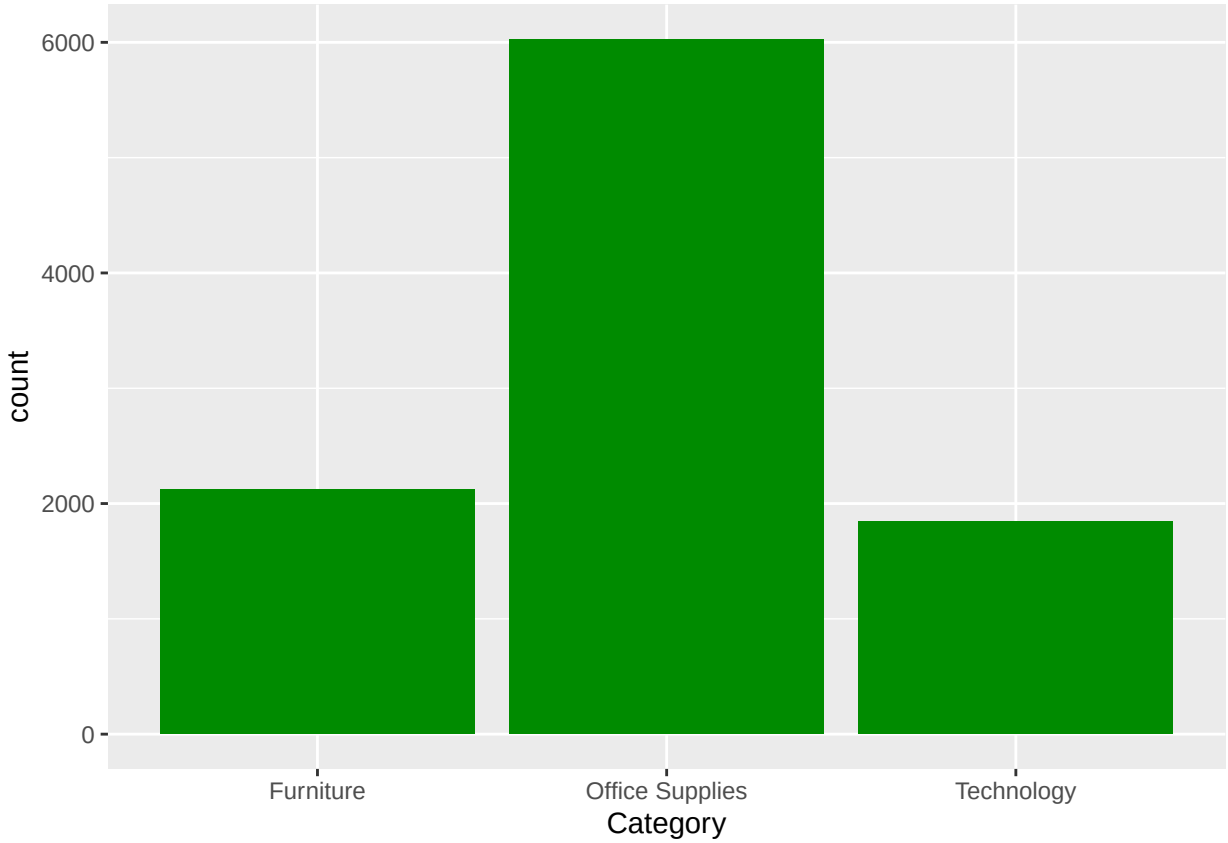
```
summary(data$Order_Date)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2014-01-03" "2015-05-23" "2016-06-26" "2016-04-30" "2017-05-14" "2017-12-30"
```

```
#[1] min "2014-01-03", max "2017-12-30"
```

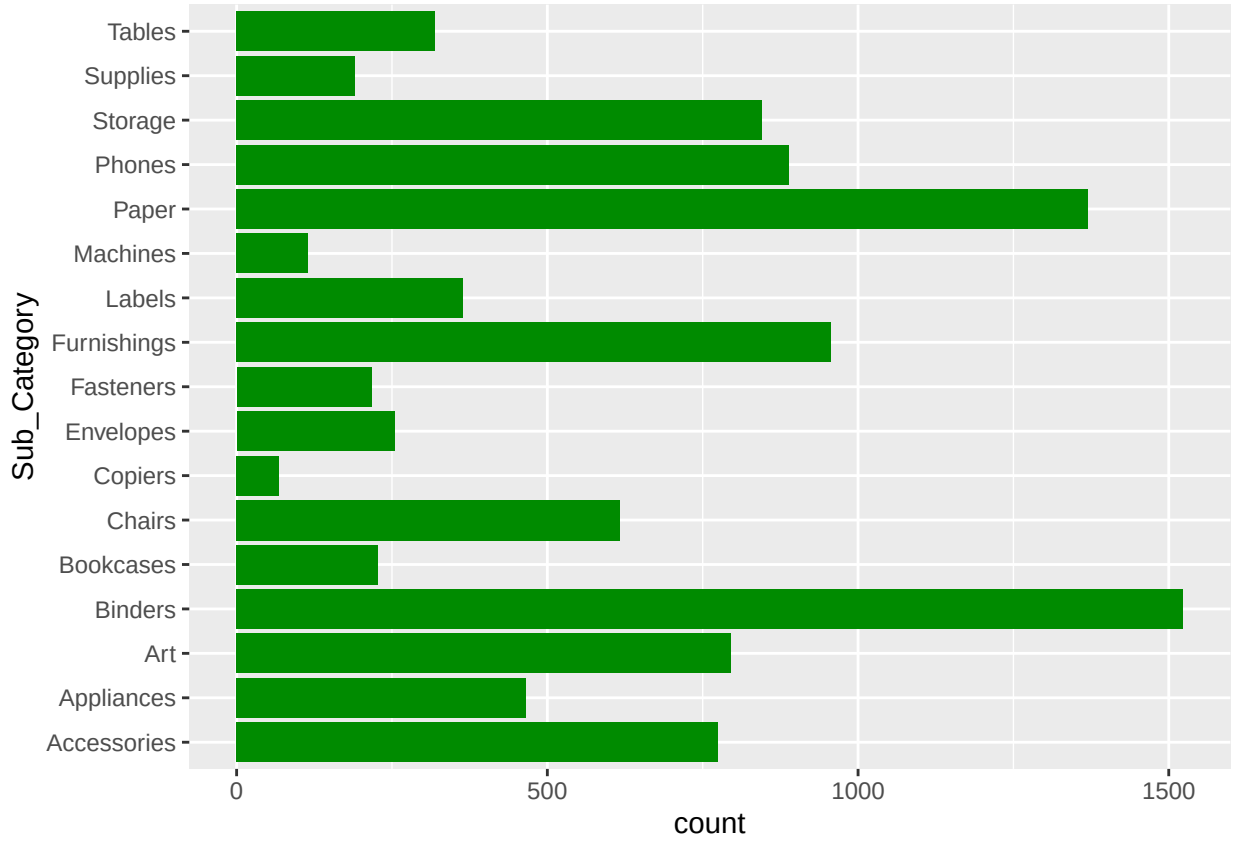
Basically this dataset covers transactions ranging from 2014-01-03 to 2017-12-30.

```
ggplot(data = data) +
  geom_bar(mapping = aes(x = Category),fill="green4")
```



Most type of products sold belong to the Office supplies category.

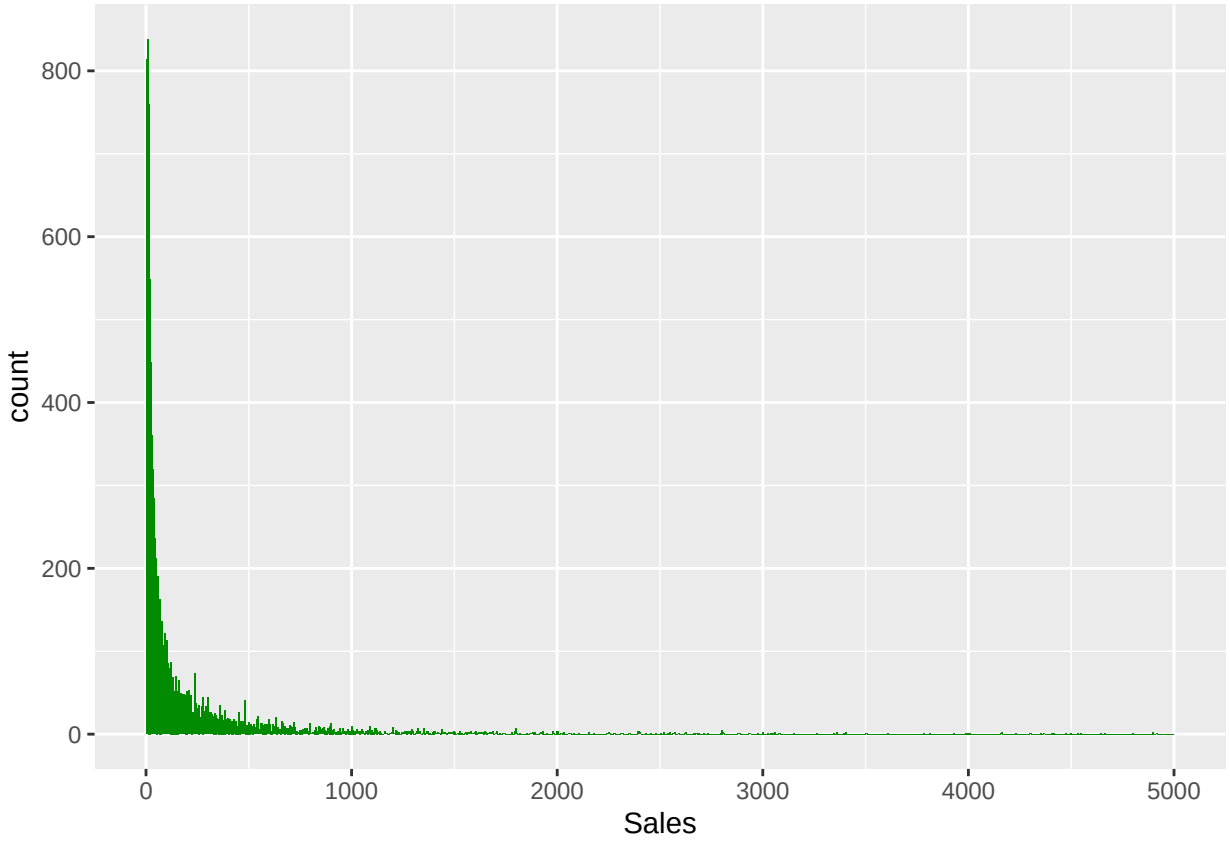
```
ggplot(data = data) +
  geom_bar(mapping = aes(y = `Sub_Category`), fill="green4")
```




```
ggplot(data = data, mapping = aes(x = Sales)) +
  xlim(0, 5000) +
  geom_histogram(binwidth = 5,fill="green4")
```

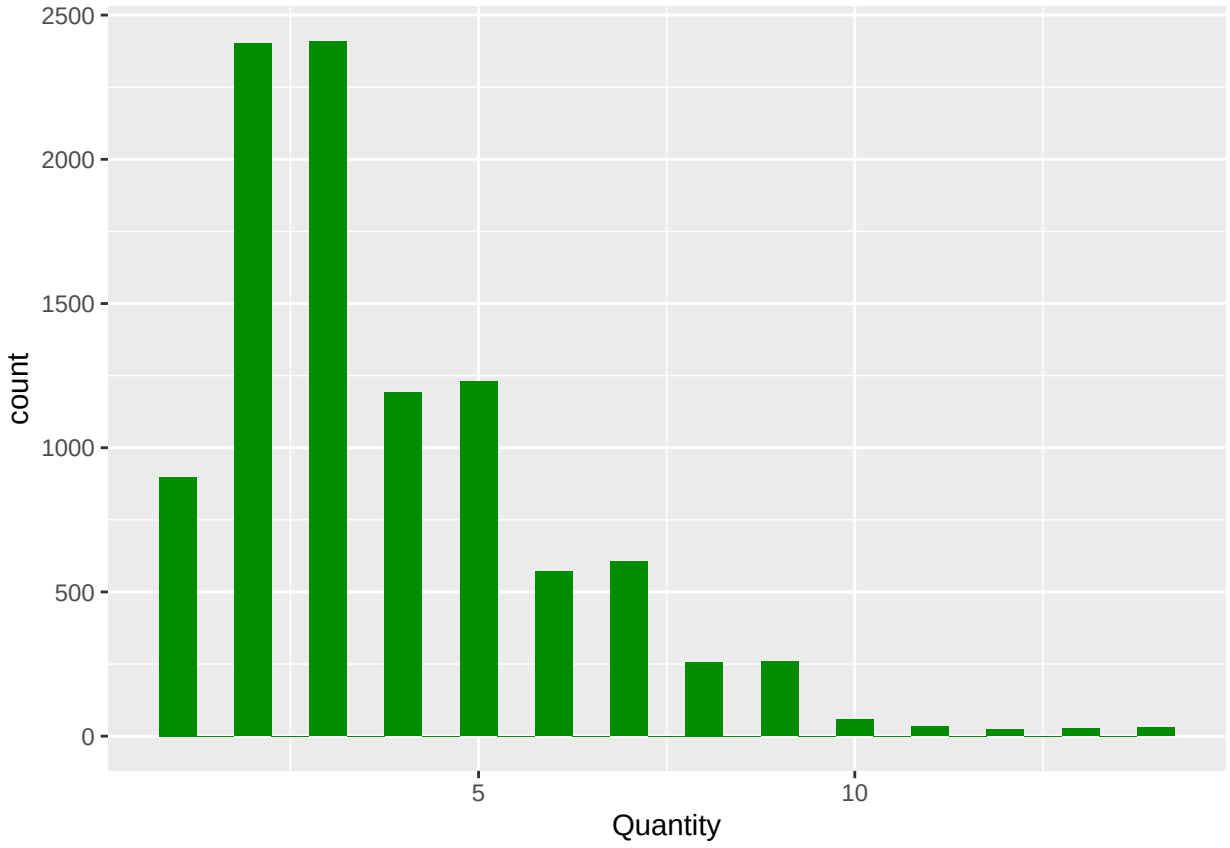
```
## Warning: Removed 19 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



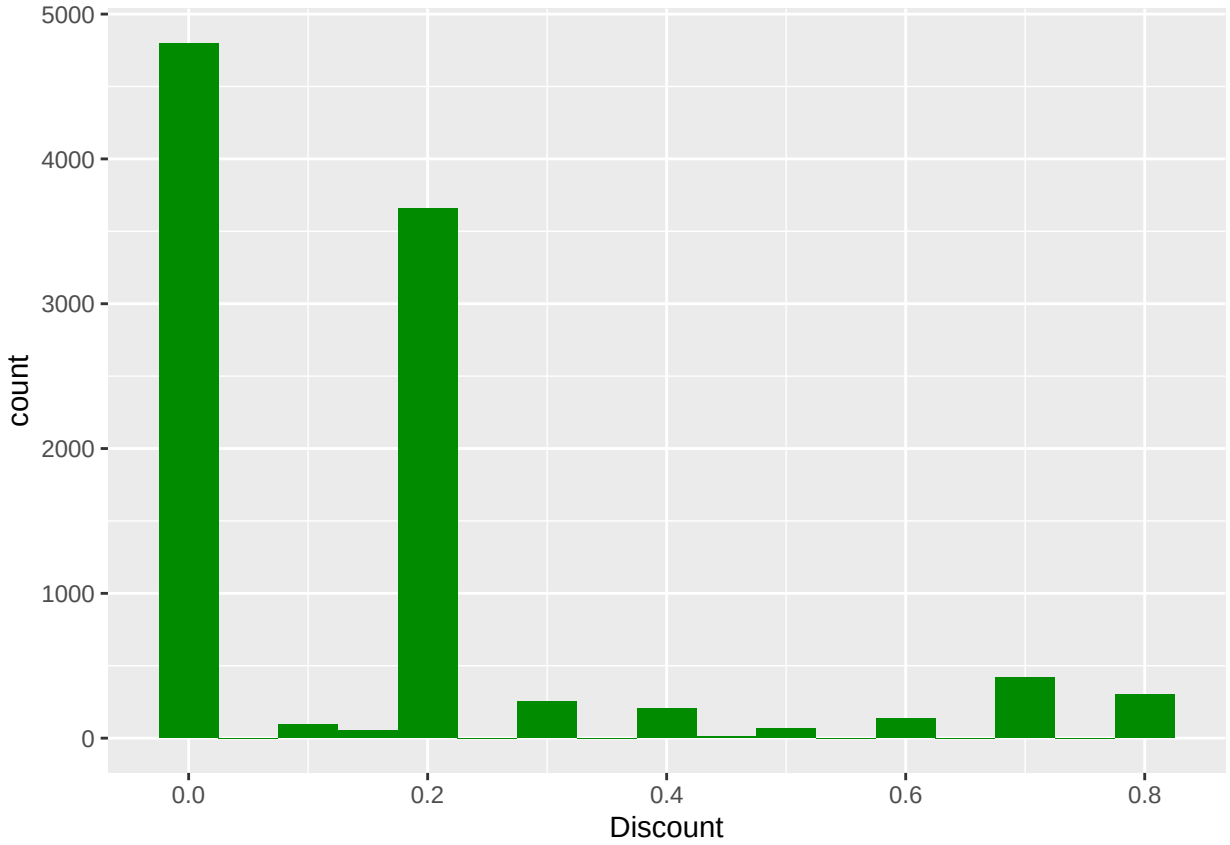
Most sales are very few items (<500).

```
ggplot(data = data, mapping = aes(x = Quantity)) +
  geom_histogram(binwidth = 0.5,fill="green4")
```




```
ggplot(data = data) +
  geom_histogram(mapping = aes(x = Discount),
                 binwidth = 0.05,
                 xlab="Discount",
                 fill="green4")
```

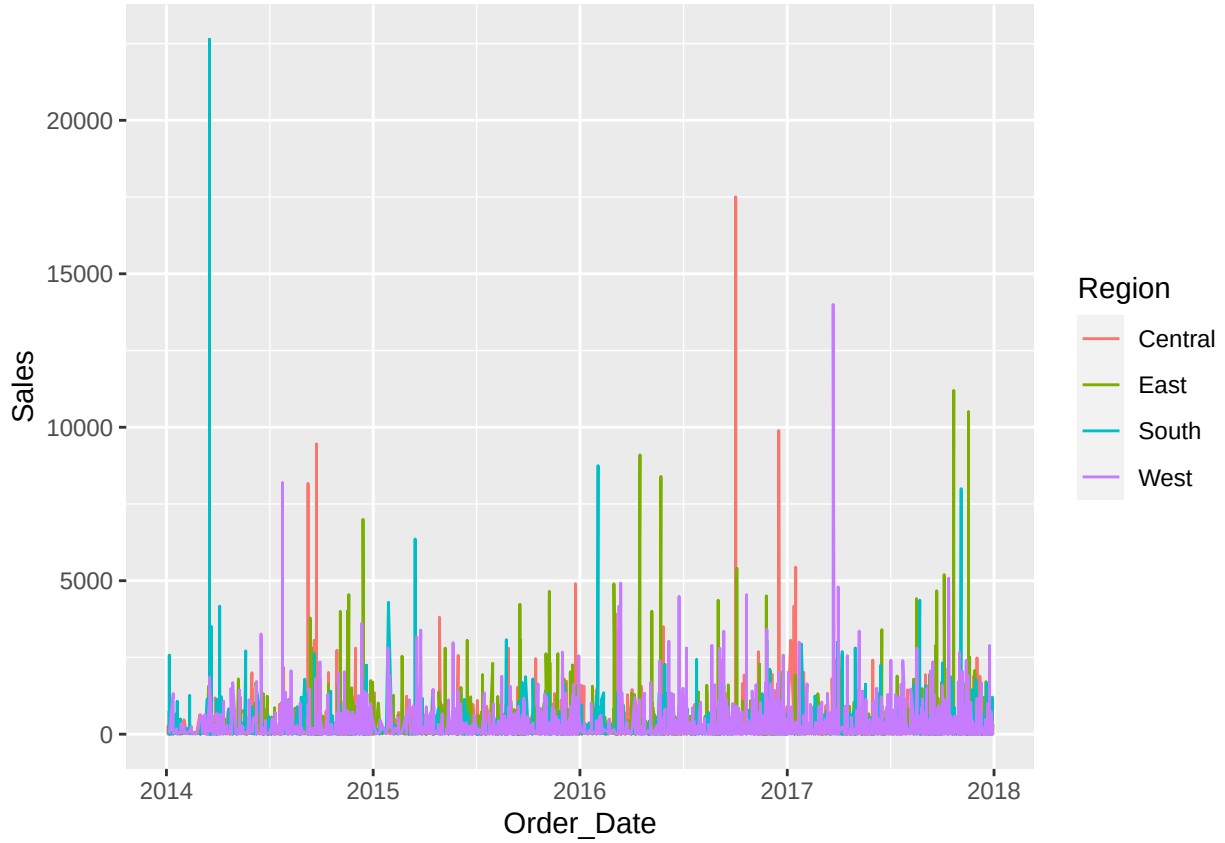
```
## Warning: Ignoring unknown parameters: xlab
```



Sales transactions mostly do not involve discounts.

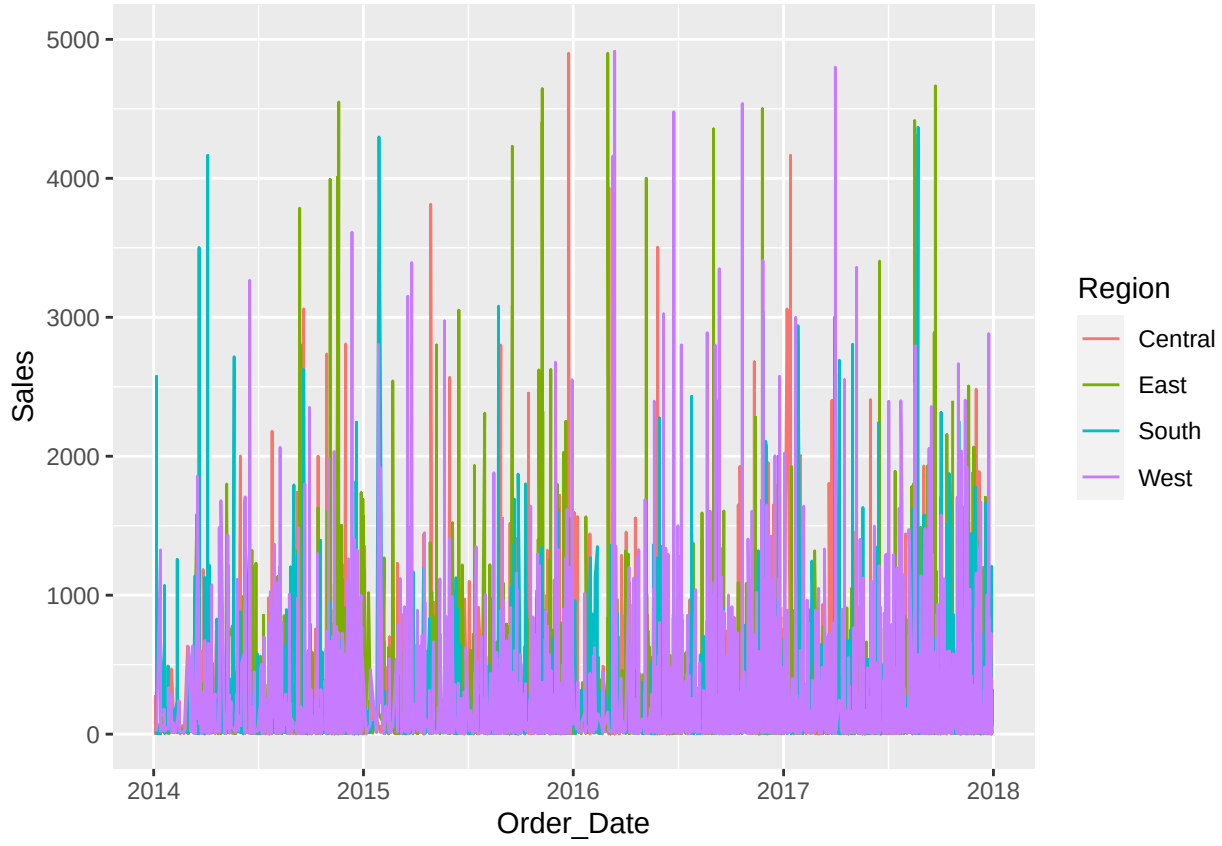
Visualise sales transactions by Region over time (order date).

```
ggplot(data, aes(Order_Date, Sales,color=Region)) +
     geom_line()
```



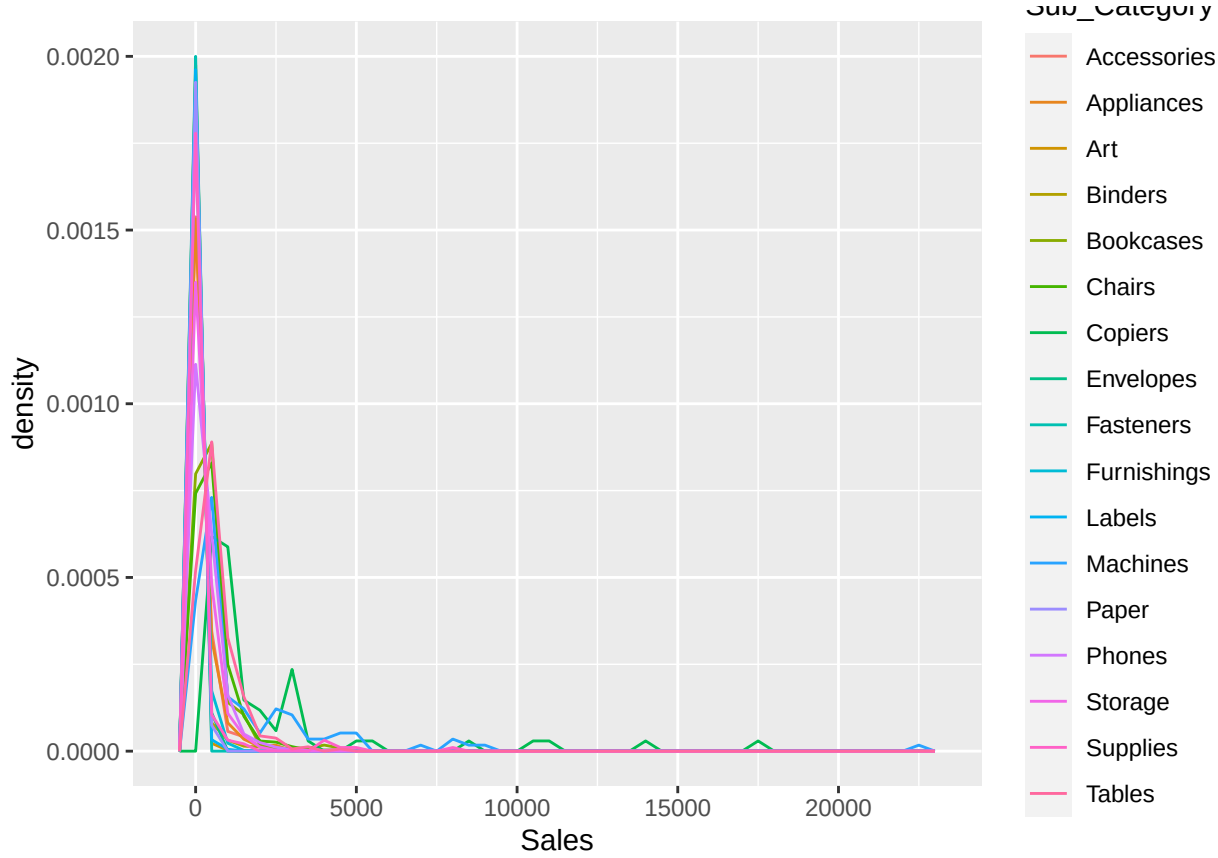
Let's zoom in a little bit - Visualise sales transactions by Region over time (order date).

```
ggplot(data, aes(Order_Date, Sales,color=Region)) +
     geom_line() +
     ylim(0,5000)
```



How does profit change with sub-category?

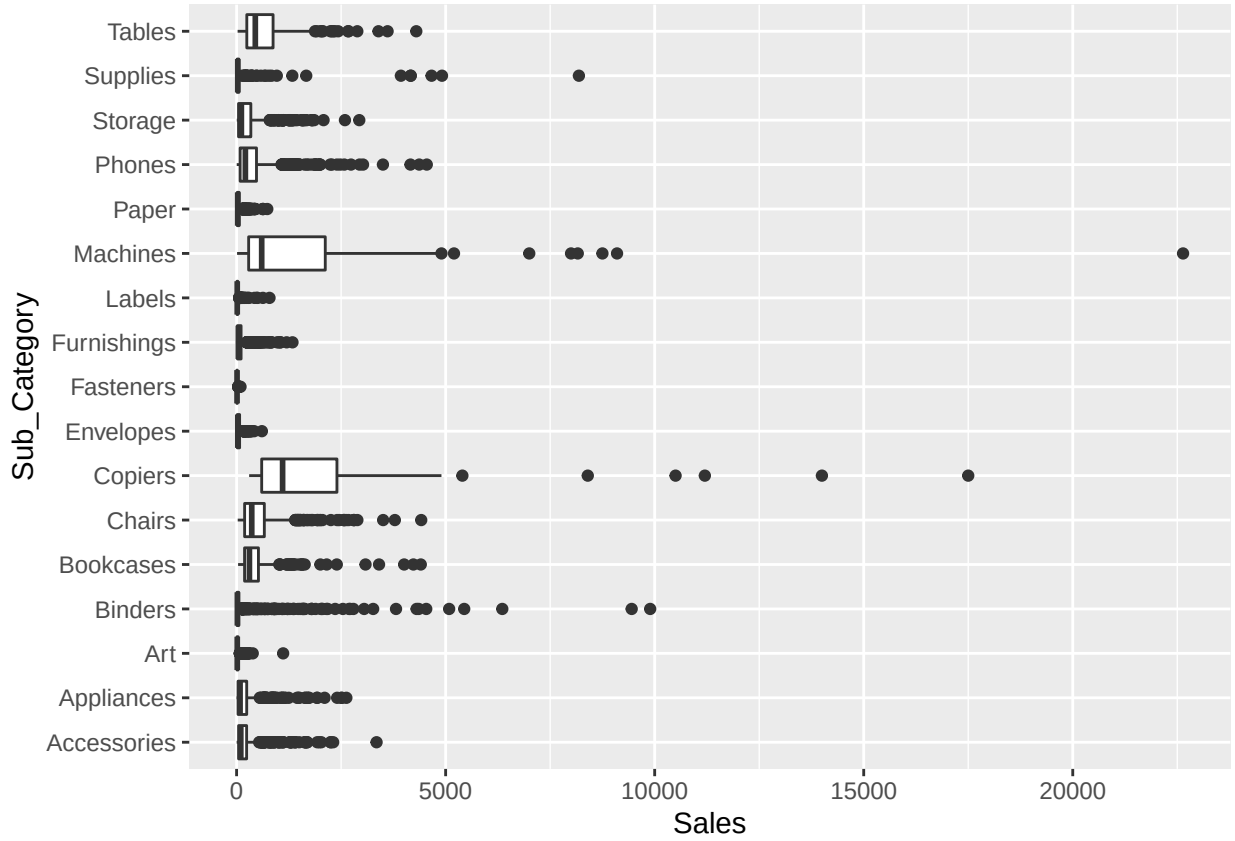
```
 #density plot where the count is standardized,area under each frequency is 1
ggplot(data = data, mapping = aes(x = Sales, y = ..density..)) +
  geom_freqpoly(mapping = aes(colour = Sub_Category), binwidth = 500)
```



It looks like some categories of items ie. supplies or accessories have negative sales values.

How does sales vary across sub category?

```
ggplot(data = data, mapping = aes(x = Sales, y = `Sub_Category` )) +
  geom_boxplot()
```



Add Pie charts! - by sub_category, region etc

Create a Month variable - to see the change of sales/profits by month?

bar charts of profits/sales by region

Output the characteristics of the orders with the highest and lowest profits/sales - e.g. what made the order? when? bought what product? in which city/state/region? Any discount?

relationship between discount & sales, discount & profits, sales & profits, and the role of region?

from someone's analysis - there is no significant change between the four discount categories when it comes to Sales

sales/profits by month, rather than by date? color by region?