Supply Chain & Logistics Analytics

Submitted by : Mohamed Yusuf S

Date of submission : Oct 2021

Objective:

To build am model to predict the correcting shipping mode

Contents

Project Objective

The objective of the report is

- To develop the best predictive model which can predict the correct shipping mode for a Supply Chain and Logistics company based on the given data,
- Compare the performance of the various models, and
- Identify the best model and which variables are a significant predictor behind this decision.

Exploratory Data Analysis

Dimensions of the data:

We have total of 7853 rows and 8 columns

```
> ##Check the dimention or shape of the data
> dim(df)
[1] 7853    8
```

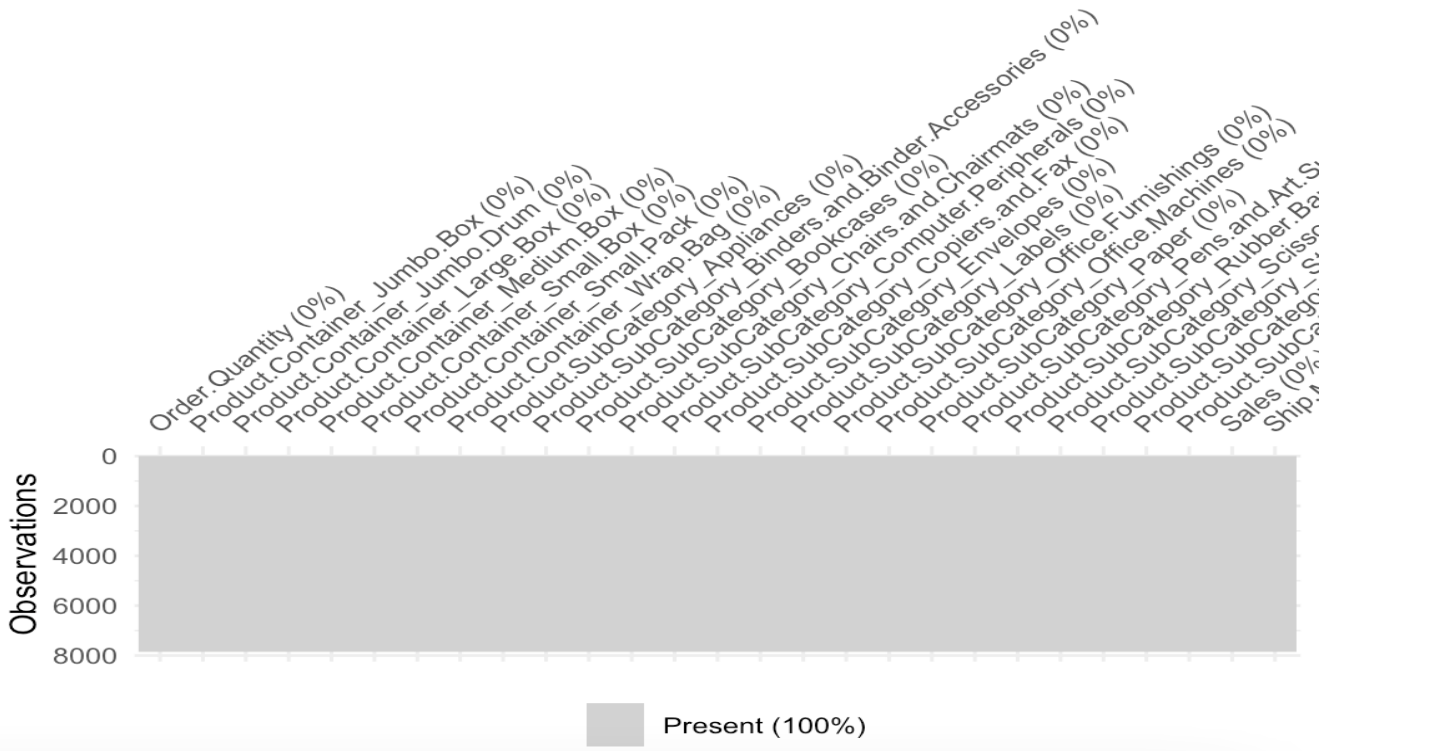
Structure of the data frame:

```
> str(df)
tibble [7,853 × 8] (S3: tbl_df/tbl/data.frame)
 $ Order Date          : chr [1:7853] "1/27/2007" "1/27/2007" "1/27/2007" "1/27/2007" ...
 $ Order ID            : num [1:7853] 24544 24544 24544 20422 55937 ...
 $ Order Quantity      : num [1:7853] 31 39 15 30 10 5 11 24 49 38 ...
 $ Product Container   : chr [1:7853] "Medium Box" "Large Box" "Jumbo Drum" "Small Pack" ...
 $ Product Name        : chr [1:7853] "Canon MP41DH Printing Calculator" "Fellowes Neat Ideas® Storage Cubes" "Global
 Stack Chair without Arms, Black" "Nu-Dell Leatherette Frames" ...
 $ Product Sub-Category: chr [1:7853] "Office Machines" "Storage & Organization" "Chairs & Chairmats" "Office Furnish
ings" ...
 $ Sales               : num [1:7853] 6567 1780 578 611 517 ...
 $ Ship Mode           : chr [1:7853] "Delivery Truck" "Regular Air" "Delivery Truck" "Regular Air" ...
```

Detecting Missing Values

No missing values found

```
> ##Identifying variable with missing values
> apply(is.na(df),2,sum)
      Order Quantity    Product Container Product Sub-Category                Sales
                   0                    0                    0                    0
         Ship Mode
                   0
```



Summary

Variables such as Ship Mode, Product Container and Product Sub-Category has been converted to factor variables.

Dropped irrelevant variables like Order date, order ID and Product name.

Summary of the data after conversion

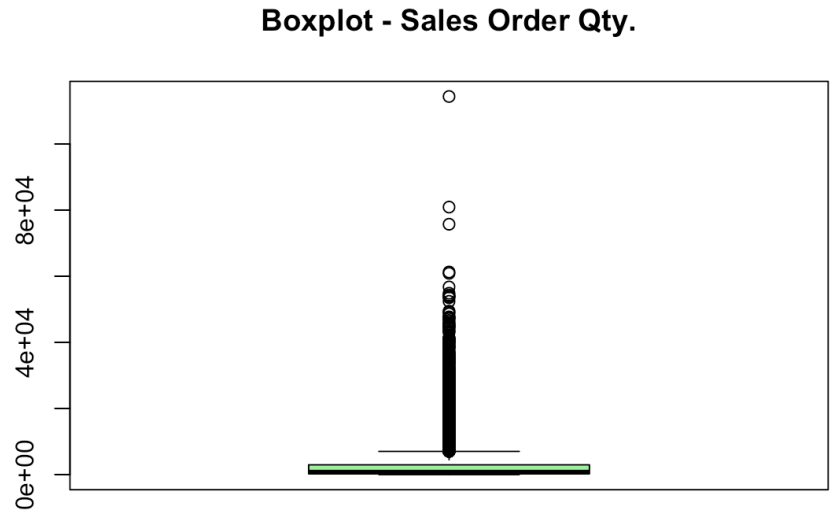
```
> summary(df)
 Order Quantity  Product Container  Product Sub-Category     Sales
 Min.   : 1.00   Length:7853        Length:7853          Min.   :     4
 1st Qu.:13.00   Class :character   Class :character     1st Qu.:   244
 Median :26.00   Mode  :character   Mode  :character     Median :   747
 Mean   :25.59                                          Mean   :  3044
 3rd Qu.:38.00                                          3rd Qu.:  2959
 Max.   :50.00                                          Max.   :114362
  Ship Mode
 Length:7853
 Class :character
 Mode  :character
```

Data Imbalance

Table of Dependent variable:

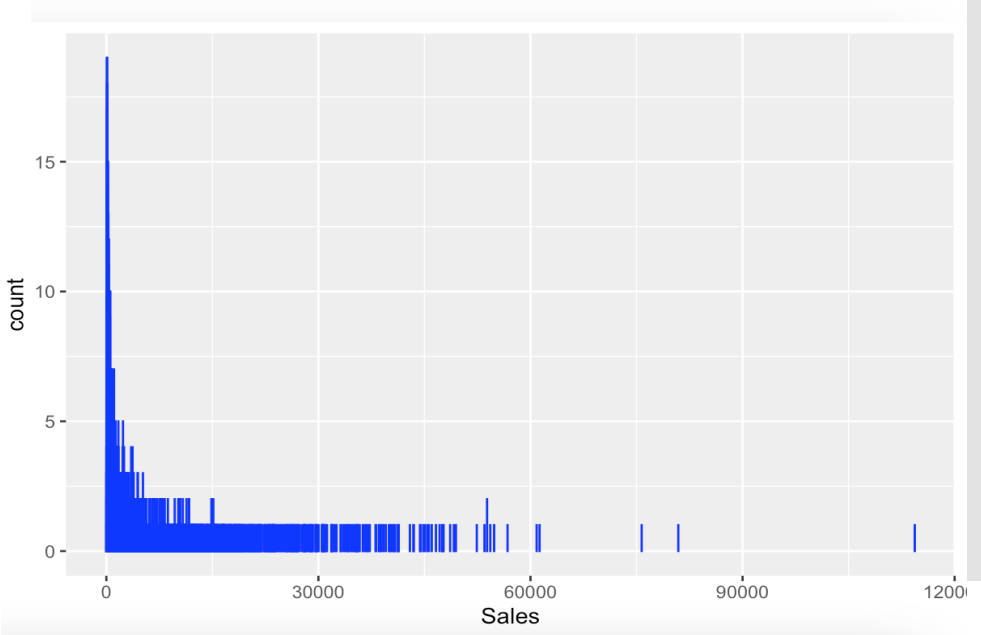
```
> # Target variable
> table(df$`Ship Mode`)

Delivery Truck     Regular Air
          1984            5869
```

The data is imbalanced.

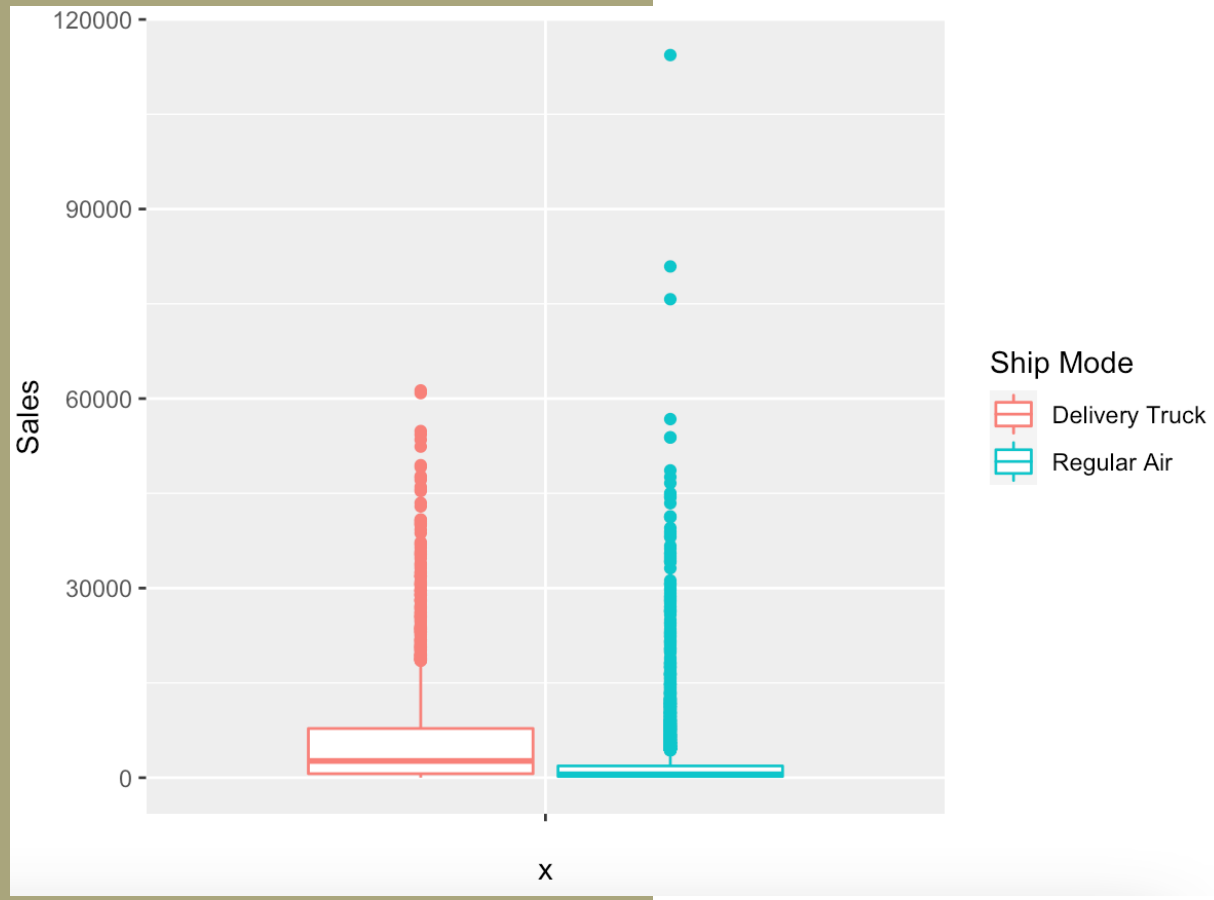


Outlier Treatment for Sales column:

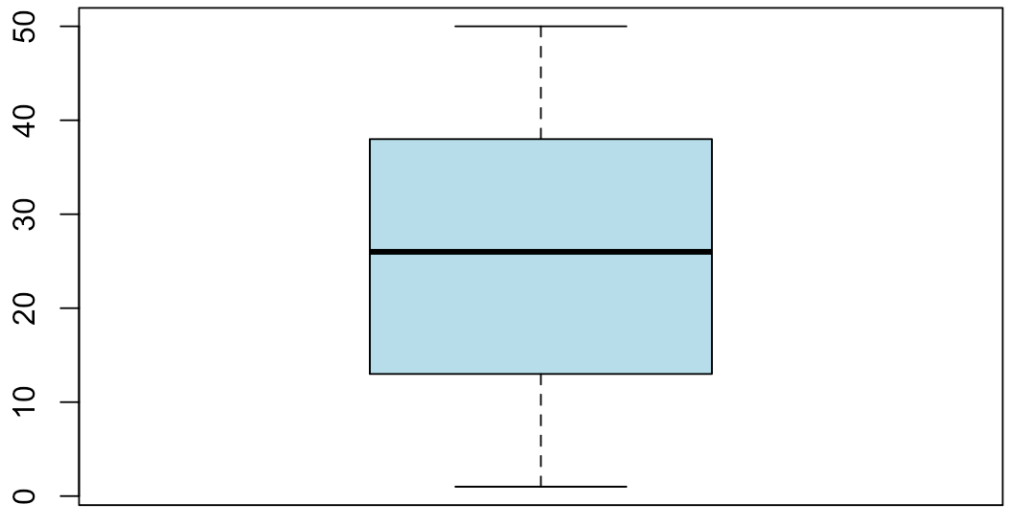
- The boxplot shows there are a good number of outliers
- We do the outlier treatment using 5th and 95th percentile values. After treatment of outliers the sales data looks like this:

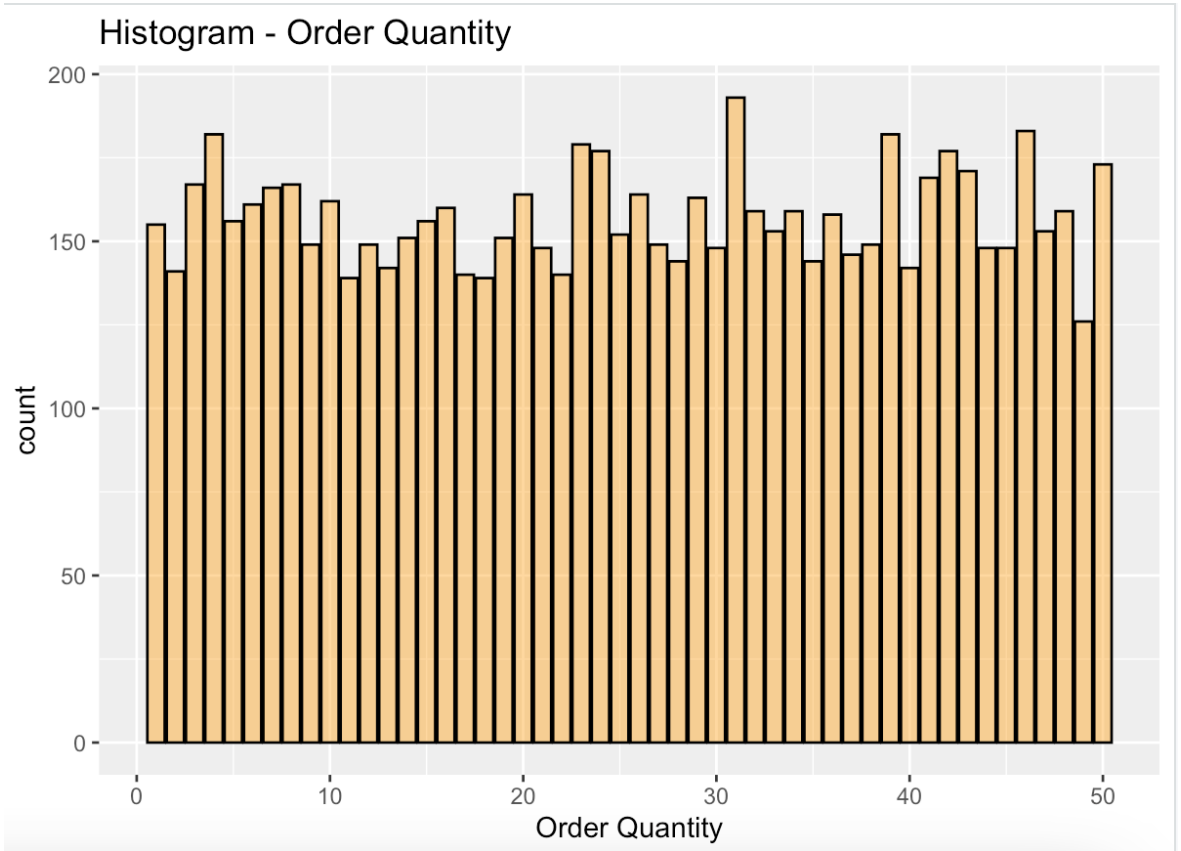


Before



After

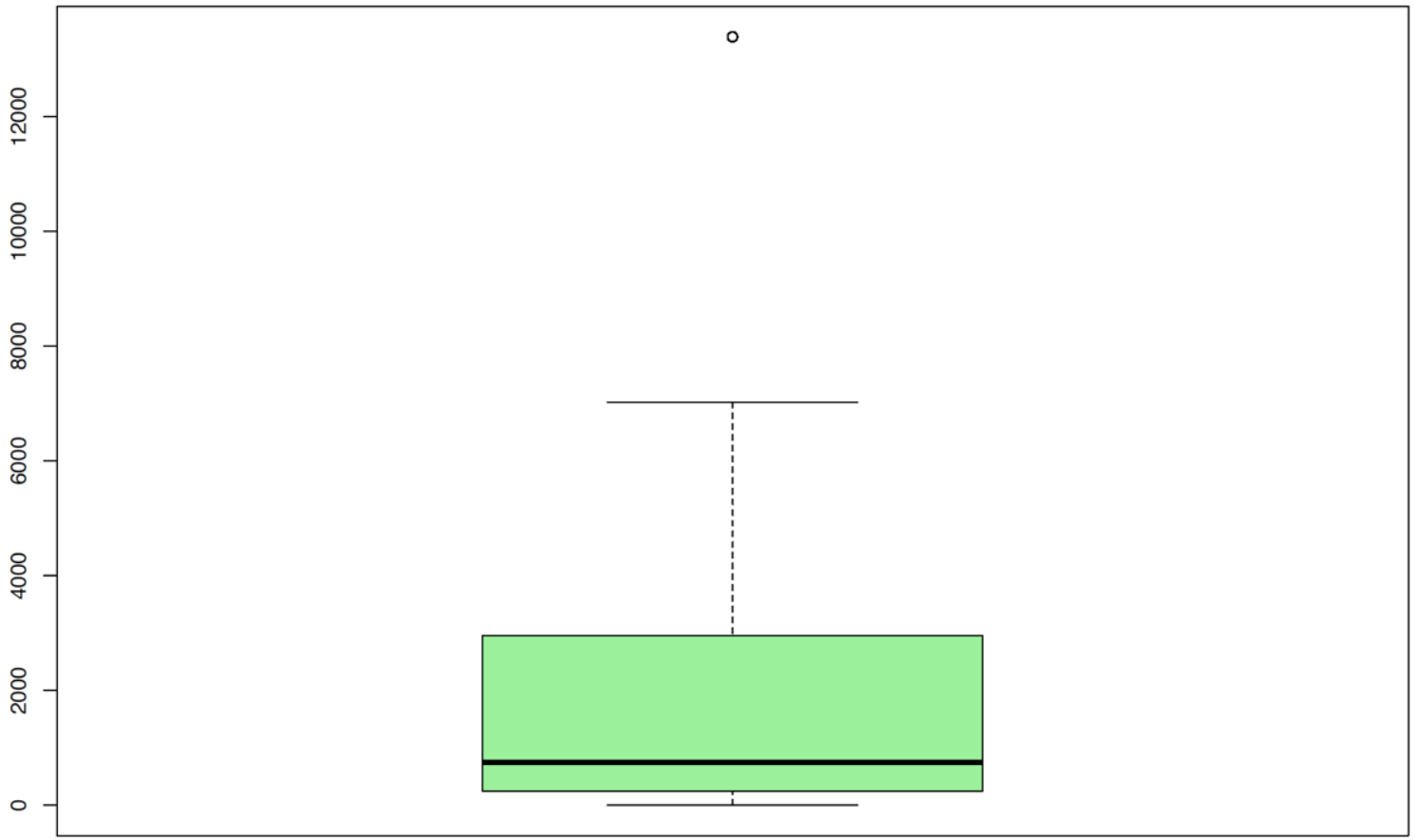


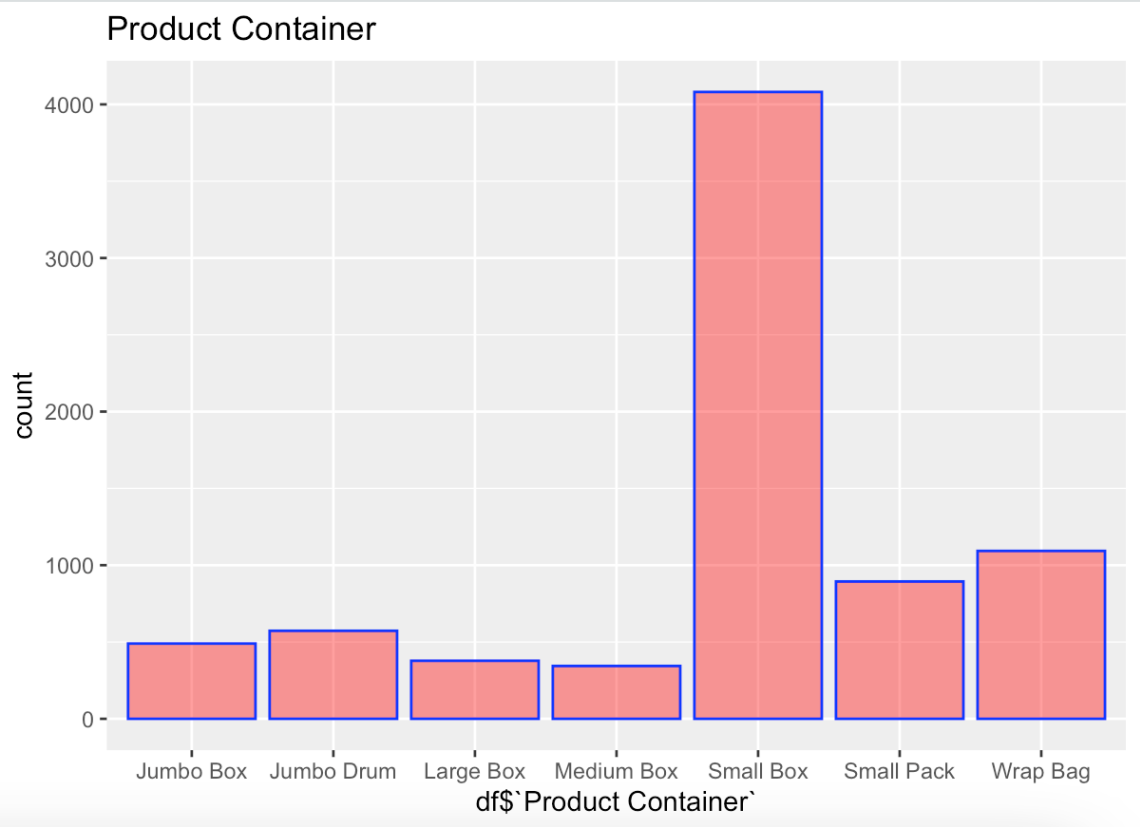


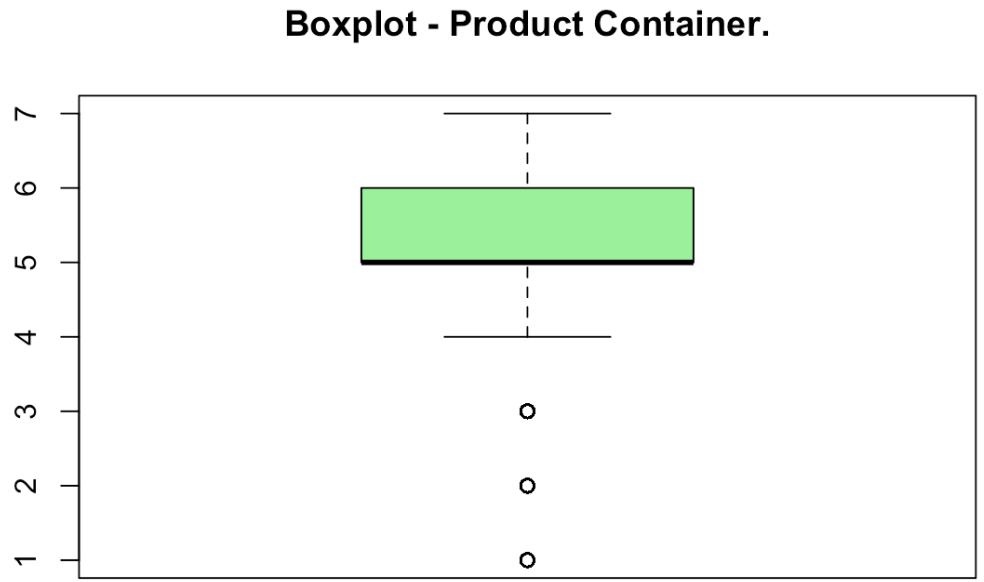
Univariate Analysis on Variable - Order Quantity

Univariate Analysis on Variable Sales

There is no much of outliers on Variable Sales after outlier treatment



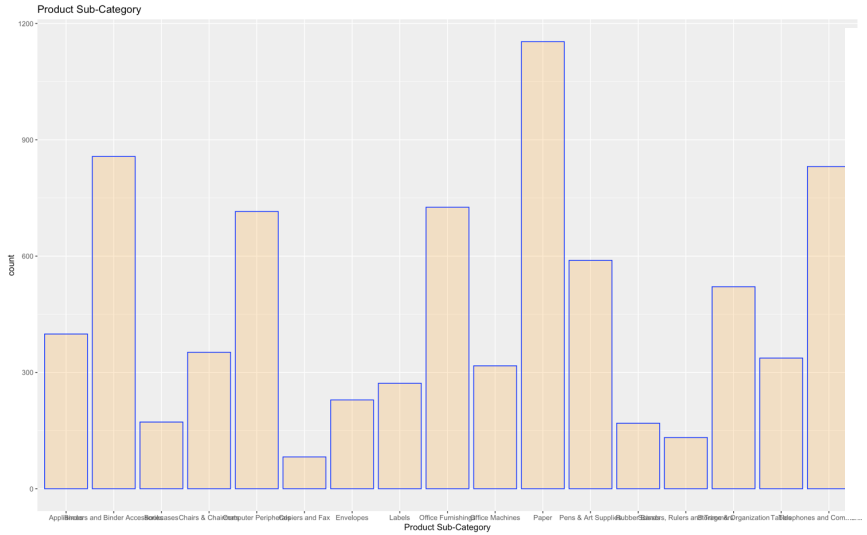


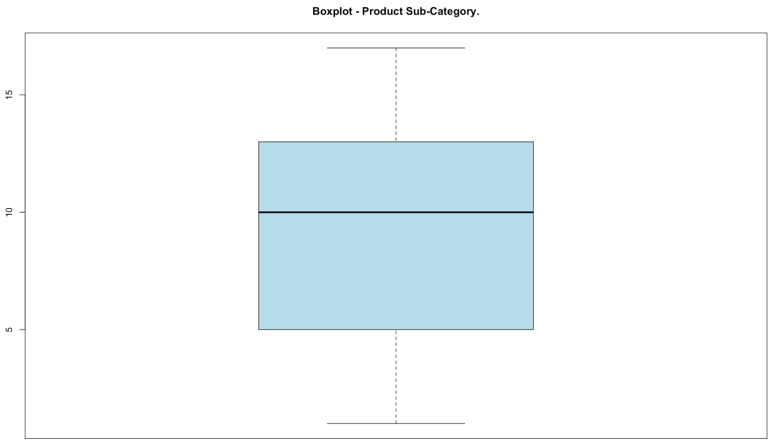



```
> summary(df$`Product Container`)
 Jumbo Box Jumbo Drum  Large Box Medium Box  Small Box Small Pack
       490        573        378        344       4081        894
 Wrap Bag
      1093
```

Small Box type have more observations compared to other Product Container types

```
> summary(df$`Product Sub-Category`)
                  Appliances Binders and Binder Accessories
                         399                             857
                  Bookcases                 Chairs & Chairmats
                         172                             352
          Computer Peripherals                   Copiers and Fax
                         715                              82
                   Envelopes                          Labels
                         229                             272
          Office Furnishings                  Office Machines
                         726                             317
                       Paper                Pens & Art Supplies
                        1153                             589
                Rubber Bands    Scissors, Rulers and Trimmers
                         169                             132
        Storage & Organization                          Tables
                         521                             337
  Telephones and Communication
                         831
```





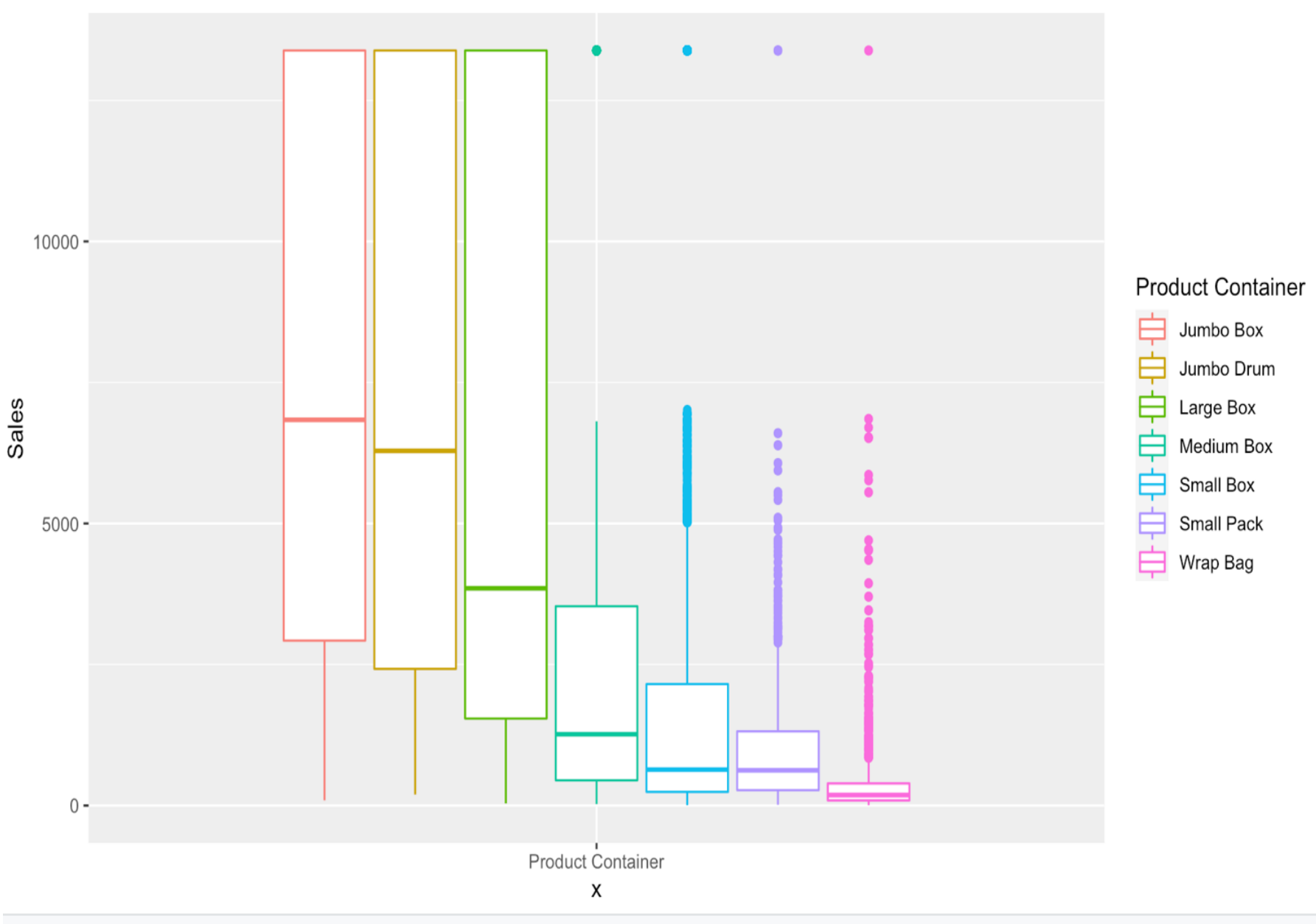
Paper, Binders and Binder Accessories and Telephone & Communication are major contributes to Product Sub-category

Bi Variate Analysis

From the above plot, Jumbo Box, Jumbo Drum and Large Box amounts to a big chunk of sales within Product Container .

Small Box, Small Pack and Wrap Bag are the Lowest contributors.

It is obvious that only small items(low cost) will be packed in small containers



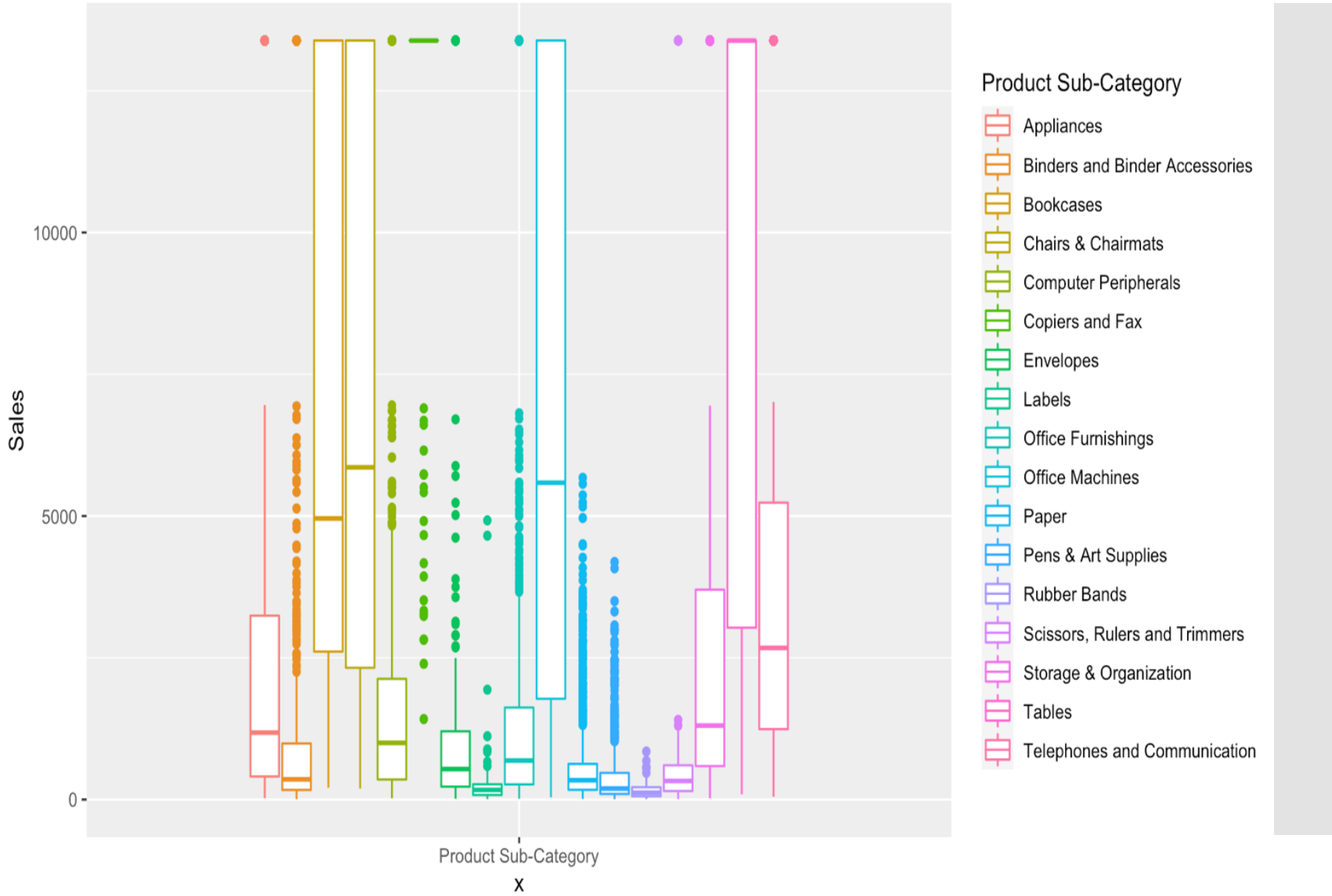
Sales Vs Product Sub-Category

Bi Variate Analysis

From the plot above, Product Sub-Categories that contributes to maximum sales are

Bookcases, Office Machines, Tables, Chairs-chair mats.

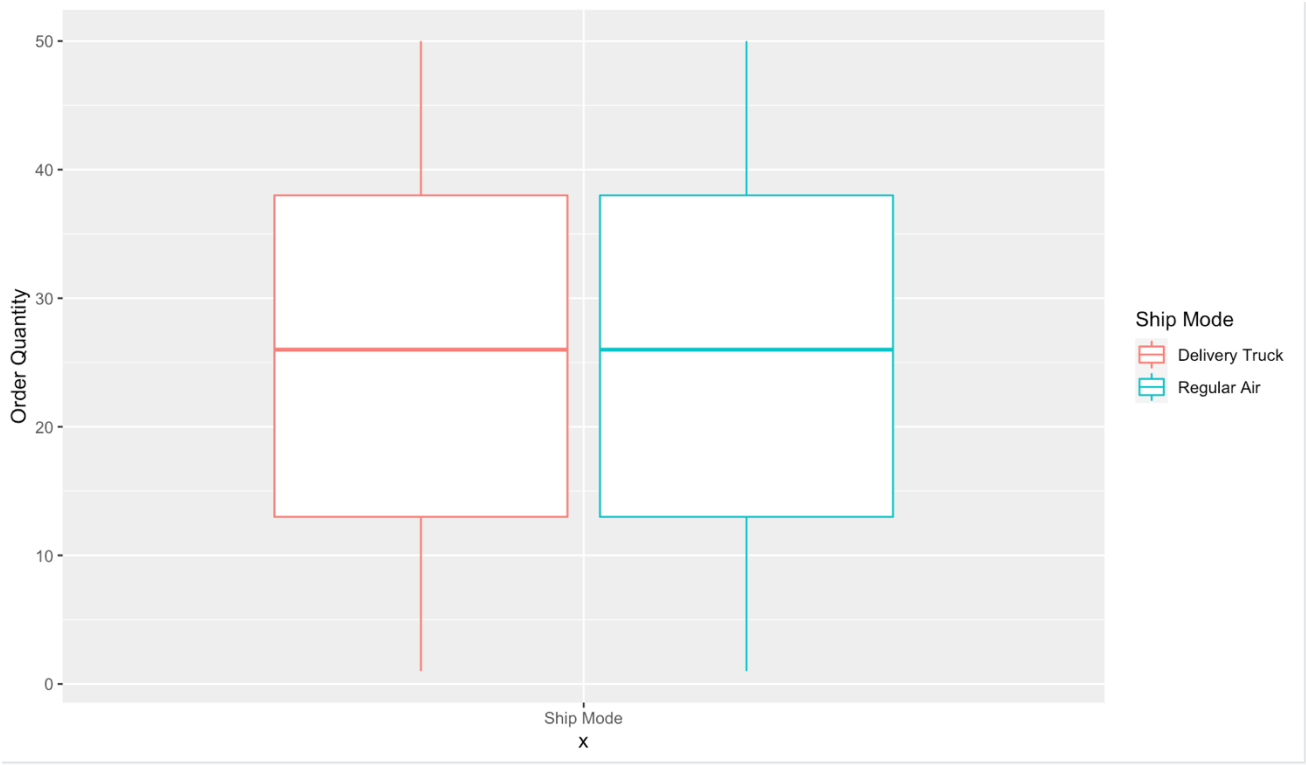
Labels, Rubber bands contribute least.



Order Quantity Vs Shipment Mode

Bivariate Analysis

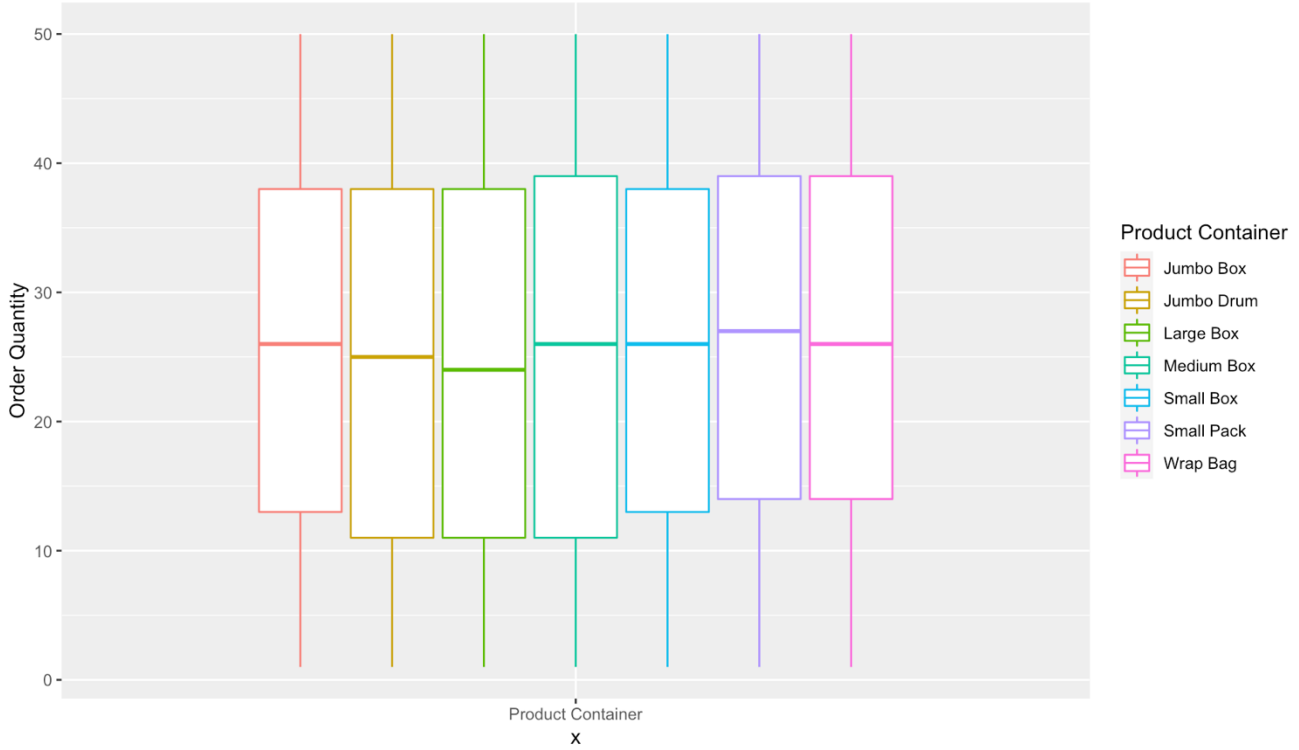
Order quantity looks much similar for both the shipment modes.



Order Quantity Vs Product Container

Bivariate Analysis

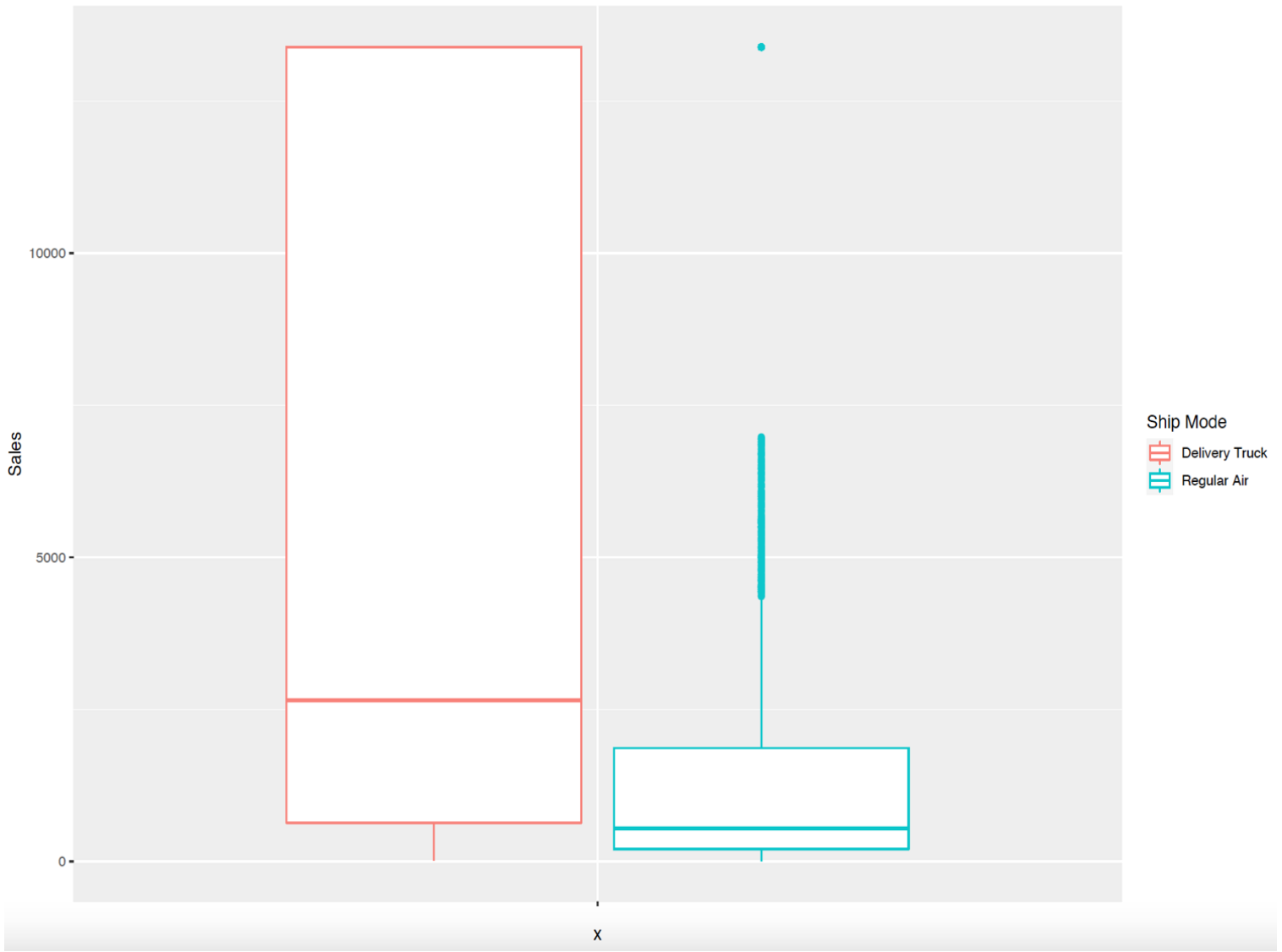
Order quantity looks very much the similar for all product containers.



Sales Vs Ship Mode

The sales were high for Delivery truck shipment mode.

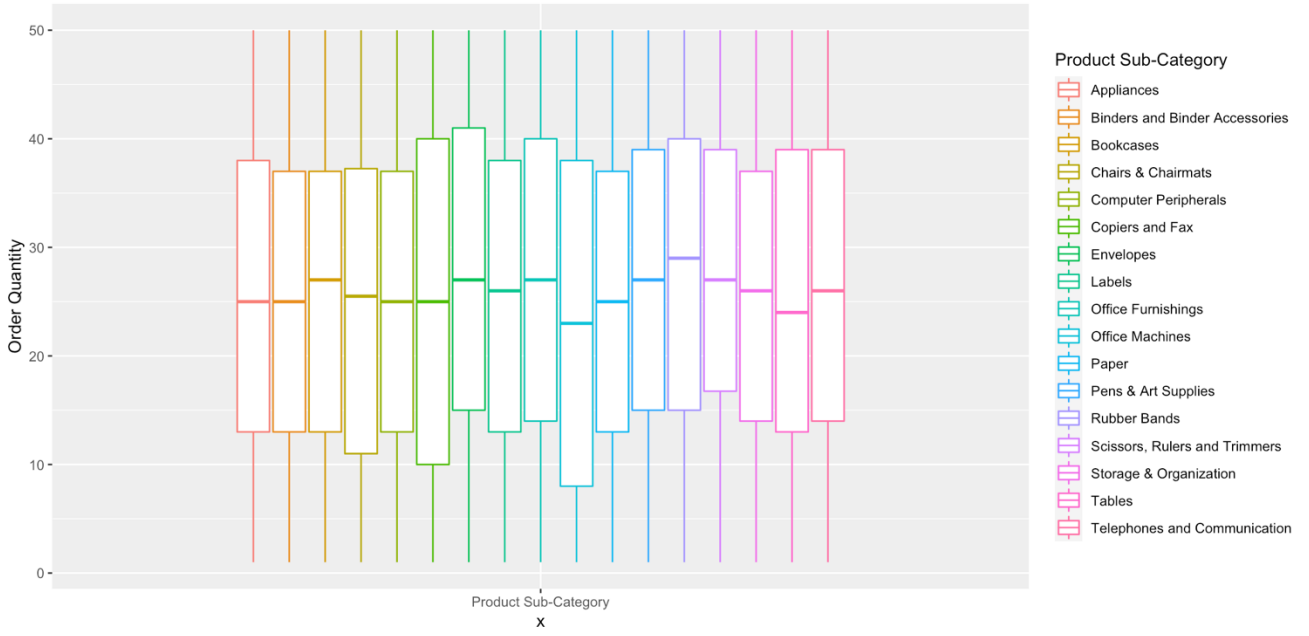
The sales looks comparatively less for Regular air shipment mode.



Bivariate Analysis

Similarly Order quantity was pretty much the same for all product sub-categories.

Order Quantity Vs Product Sub-Category

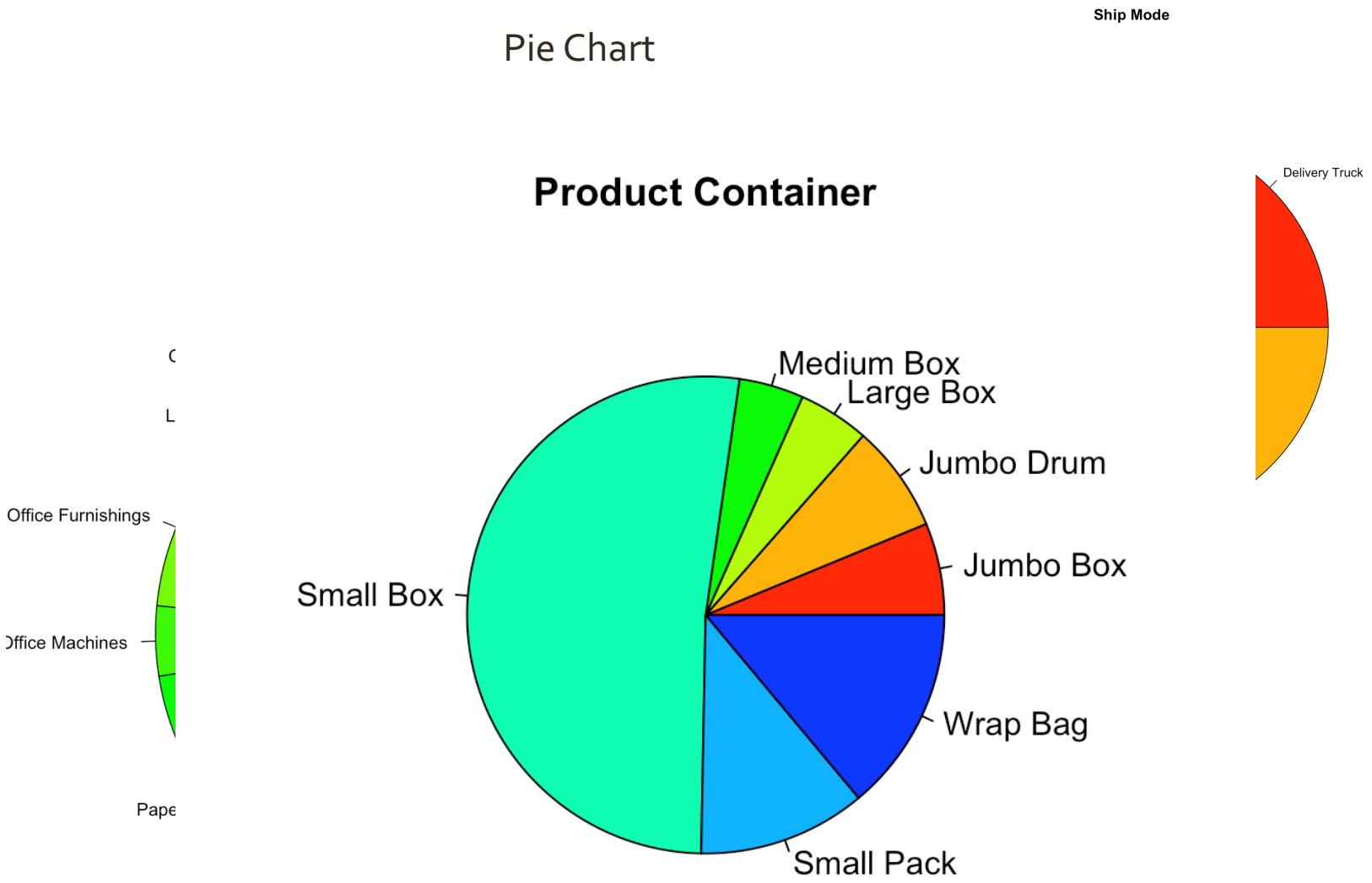


Pie Chart

Paper sales is the highest in product sub-category

Regular Air is the most preferred shipment mode.

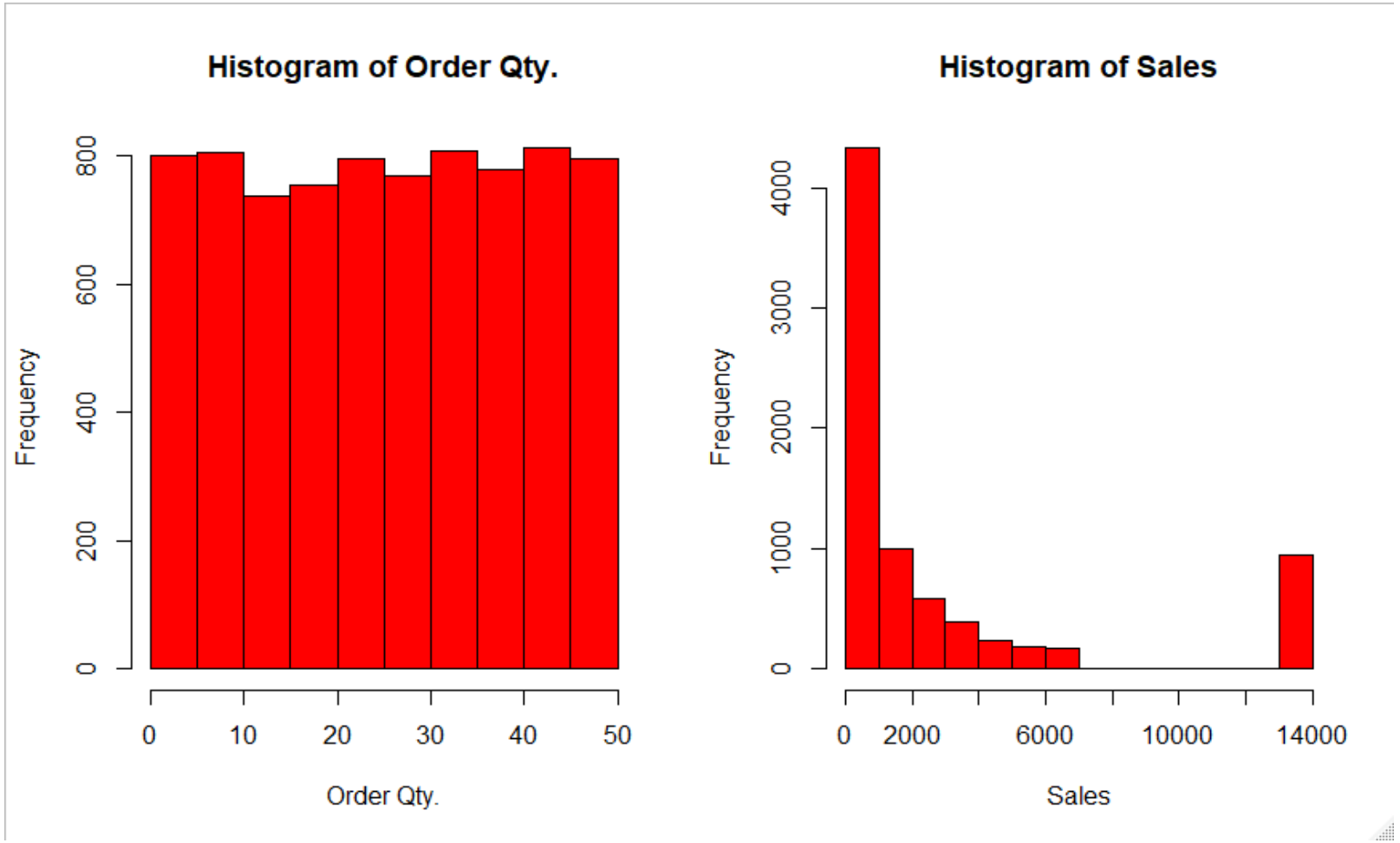
Small box is highly used as Product Container.

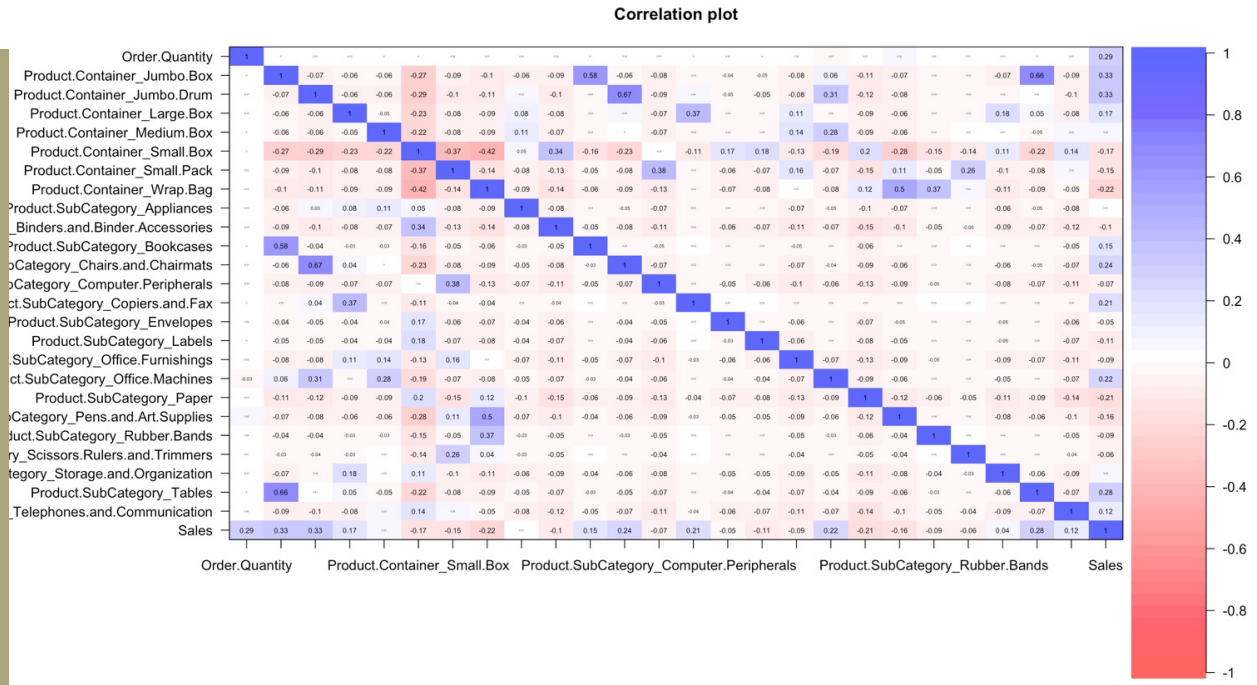


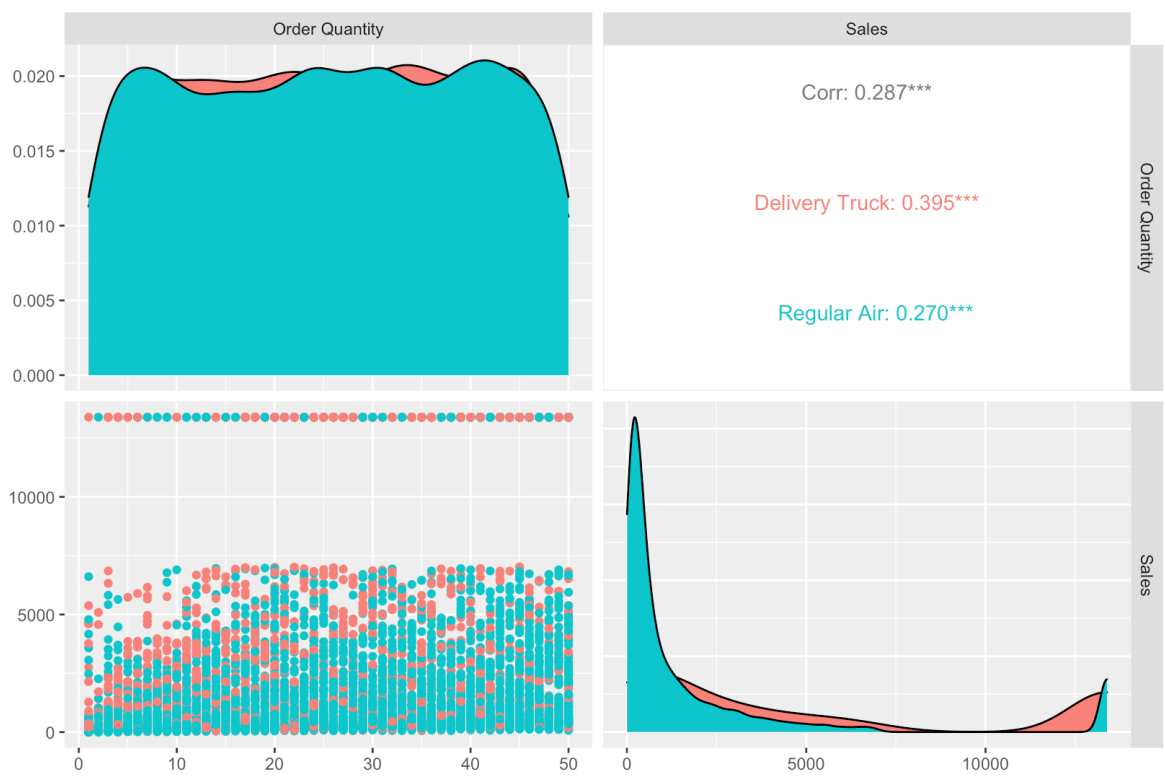
Histogram

Frequency of Order Quantity distribution very similar to order quantity.

Sales has a right skewed distribution.




Correlation plot
Order.Quantity
Product.Container_Jumbo.Box
Product.Container_Jumbo.Drum
Product.Container_Large.Box
Product.Container_Medium.Box
Product.Container_Small.Box
Product.Container_Small.Pack
Product.Container_Wrap.Bag
Product.SubCategory_Appliances
_Binders.and.Binder.Accessories
Product.SubCategory_Bookcases
Category_Chairs.and.Chairmats
Category_Computer.Peripherals
ct.SubCategory_Copiers.and.Fax
Product.SubCategory_Envelopes
Product.SubCategory_Labels
.SubCategory_Office.Furnishings
ct.SubCategory_Office.Machines
Product.SubCategory_Paper
Category_Pens.and.Art.Supplies
duct.SubCategory_Rubber.Bands
ry_Scissors.Rulers.and.Trimmers
tegory_Storage.and.Organization
Product.SubCategory_Tables
_Telephones.and.Communication
Sales
Order.Quantity
Product.Container_Small.Box
Product.SubCategory_Computer.Peripherals
Product.SubCategory_Rubber.Bands
Sales
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Order Quantity
Sales
Corr: 0.287***
Delivery Truck: 0.395***
Regular Air: 0.270***
Order Quantity
Sales
0.020
0.015
0.010
0.005
0.000
10000
5000
0
0
10
20
30
40
50
0
5000
10000

Feature Engineering

One hot encoding

- The given dataset contains two independent variables and 1 target variable which are factors. So, we have done one-hot encoding on independent variables.
- For target variable we have done encoding where 1 stand for Regular Air, 2 for Delivery Truck.
- One hot encoding is done because it allows representation of categorical data more expressive. Many machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers. This is required for both input and output variables that are categorical.

Structure of the data after encoding

Except for Order Quantity and Sales, other variables have been factored for better modeling

```
> str(df)
'data.frame':   7853 obs. of  27 variables:
 $ Order.Quantity                                    : num  31 39 15 30 10 5 11 24 49 38 ...
 $ Product.Container_Jumbo.Box                       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
 $ Product.Container_Jumbo.Drum                      : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
 $ Product.Container_Large.Box                       : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
 $ Product.Container_Medium.Box                      : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
 $ Product.Container_Small.Box                       : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 2 1 1 2 ...
 $ Product.Container_Small.Pack                      : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
 $ Product.Container_Wrap.Bag                        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Appliances                    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Binders.and.Binder.Accessories: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Bookcases                     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Chairs.and.Chairmats          : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Computer.Peripherals          : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
 $ Product.SubCategory_Copiers.and.Fax               : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Envelopes                     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Labels                        : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
 $ Product.SubCategory_Office.Furnishings            : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 2 1 ...
 $ Product.SubCategory_Office.Machines               : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Paper                         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
 $ Product.SubCategory_Pens.and.Art.Supplies         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Rubber.Bands                  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Scissors.Rulers.and.Trimmers  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Storage.and.Organization      : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
 $ Product.SubCategory_Tables                        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
 $ Product.SubCategory_Telephones.and.Communication  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
 $ Sales                                             : num  6567 1780 578 611 517 ...
 $ Ship.Mode                                         : Factor w/ 2 levels "1","2": 2 1 2 1 1 1 1 2 1 1 ...
```

We will be using the above factored data to model.

Data Imbalance

There is a high imbalance in test and train set of target variable

After splitting the data into test and train, we see that data is highly imbalanced:

```
> #Compare response variable levels in train & test
> round(prop.table(table(sdata$Ship.Mode))*100,digits = 1)

   1    2
74.7 25.3
> round(prop.table(table(train1$`Ship.Mode`))*100, digits = 1)

   1    2
75.7 24.3
> round(prop.table(table(test1$`Ship.Mode`))*100, digits = 1)

   1    2
72.5 27.5
```

Data Imbalance

Train dataset is highly imbalanced with higher positive class

```
> # Do you think data is imbalanced? Use appropriate measures to balance data.
> trainTask1
Supervised task: train1
Type: classif
Target: Ship.Mode
Observations: 5497
Features:
   numerics     factors     ordered functionals
          2          24           0           0
Missings: FALSE
Has weights: FALSE
Has blocking: FALSE
Has coordinates: FALSE
Classes: 2
   1    2
4161 1336
Positive class: 1
> 1336/4161
[1] 0.3210767
```

After using SMOTE technique, the data is well balanced for purpose of unbiased modelling

We did an under sample of the data which is much better than previous one

```
> trainTaskBal1 <- smote(trainTask1, rate=3, nn=5) # SMOTE with 5 nearest neighbours, by level of 6
> trainTaskBal1 #The number of minority class records have been increased by factor of 4
Supervised task: train1
Type: classif
Target: Ship.Mode
Observations: 8169
Features:
   numerics     factors     ordered functionals
          2          24           0           0
Missings: FALSE
Has weights: FALSE
Has blocking: FALSE
Has coordinates: FALSE
Classes: 2
   1    2
4161 4008
Positive class: 1
> 4008/4161
[1] 0.96323
```

Modelling – Logistic Regression

After cross validation, the accuracy has improved. Product Container Jumbo Box and Jumbo Drum have higher co-efficient

```
> #cross validation accuracy
> cv.logistic$aggr
acc.test.mean
      0.77231
> cv.logistic$measures.test
  iter       acc
1    1 0.7686375
2    2 0.7715755
3    3 0.7767169
> #train model
> cvmodel <- train(logistic.learner,trainTaskBal1)
> getLearnerModel(cvmodel)

Call:  stats::glm(formula = f, family = "binomial", data = getTaskData(.task,
    .subset), weights = .weights, model = FALSE)

Coefficients:
                                        (Intercept)
                                         -0.6720358
                                       Order.Quantity
                                          0.0020397
                           Product.Container_Jumbo.Box1
                                         19.5675361
                          Product.Container_Jumbo.Drum1
                                         19.4890146
                           Product.Container_Large.Box1
                                          0.3124444
                          Product.Container_Medium.Box1
                                          0.3106474
                           Product.Container_Small.Box1
                                          0.1147272
```

Summary of the model:

```
                                    Product.Container_Small.Pack1
                                                        -0.2018352
                                      Product.Container_Wrap.Bag1
                                                               NA
                                 Product.SubCategory_Appliances1
                                                        0.1333180
Product.SubCategory_Binders.and.Binder.Accessories1
                                                        -0.4986253
                                  Product.SubCategory_Bookcases1
                                                        -0.0976047
                       Product.SubCategory_Chairs.and.Chairmats1
                                                        0.1105467
                   Product.SubCategory_Computer.Peripherals1
                                                        -0.0368539
                        Product.SubCategory_Copiers.and.Fax1
                                                        0.3806067
                                Product.SubCategory_Envelopes1
                                                        -0.1783011
                                   Product.SubCategory_Labels1
                                                        -0.4443026
                     Product.SubCategory_Office.Furnishings1
                                                        -0.2114337
                        Product.SubCategory_Office.Machines1
                                                        -0.1090866
                                  Product.SubCategory_Paper1
                                                        -0.2588195
                   Product.SubCategory_Pens.and.Art.Supplies1
                                                        -0.2253505
                                Product.SubCategory_Rubber.Bands1
                                                        -0.5288416
  Product.SubCategory_Scissors.Rulers.and.Trimmers1
                                                        -0.0368539
                        Product.SubCategory_Copiers.and.Fax1
                                                        0.3806067
                                Product.SubCategory_Envelopes1
                                                        -0.1783011
                                   Product.SubCategory_Labels1
                                                        -0.4443026
                     Product.SubCategory_Office.Furnishings1
                                                        -0.2114337
                        Product.SubCategory_Office.Machines1
                                                        -0.1090866
                                  Product.SubCategory_Paper1
                                                        -0.2588195
                   Product.SubCategory_Pens.and.Art.Supplies1
                                                        -0.2253505
                                Product.SubCategory_Rubber.Bands1
                                                        -0.5288416
  Product.SubCategory_Scissors.Rulers.and.Trimmers1
                                                        -0.0395882
      Product.SubCategory_Storage.and.Organization1
                                                        -0.2809000
                                   Product.SubCategory_Tables1
                                                        0.0076346
  Product.SubCategory_Telephones.and.Communication1
                                                               NA
                                                            Sales
                                                        -0.0000411

Degrees of Freedom: 8168 Total (i.e. Null);  8144 Residual
Null Deviance:      11320
Residual Deviance: 7370         AIC: 7420
```

Confusion matrix – Logistics Regression

Model has performed well good precision, Recall and F1 scores.

Logistic regression has given good results with accuracy of 0.87.

```
    predicted
true 1         2
   1 1708      0      tpr: 1    fnr: 0
   2 298       350    fpr: 0.46 tnr: 0.54
     ppv: 0.85 for: 0 lrp: 2.17 acc: 0.87
     fdr: 0.15 npv: 1 lrm: 0    dor: Inf


Abbreviations:
tpr - True positive rate (Sensitivity, Recall)
fpr - False positive rate (Fall-out)
fnr - False negative rate (Miss rate)
tnr - True negative rate (Specificity)
ppv - Positive predictive value (Precision)
for - False omission rate
lrp - Positive likelihood ratio (LR+)
fdr - False discovery rate
npv - Negative predictive value
acc - Accuracy
lrm - Negative likelihood ratio (LR-)
dor - Diagnostic odds ratio
```

Modelling - Naïve Bayes

Confusion Matrix

The Naïve bayes model has got accuracy of 82% which low compared to Logistic Regression model. The miss rate is little higher comparatively

```
> nb.model1 = train(nb.learner, trainTaskBal1)
> nb.predict1 = predict(nb.model1, testTask1)
> nb.cm1 = calculateROCMeasures(nb.predict1)
> nb.cm1
    predicted
true 1         2
   1 1561      147       tpr: 0.91 fnr: 0.09
   2 278       370       fpr: 0.43 tnr: 0.57
     ppv: 0.85 for: 0.28 lrp: 2.13 acc: 0.82
     fdr: 0.15 npv: 0.72 lrm: 0.15 dor: 14.13


Abbreviations:
tpr - True positive rate (Sensitivity, Recall)
fpr - False positive rate (Fall-out)
fnr - False negative rate (Miss rate)
tnr - True negative rate (Specificity)
ppv - Positive predictive value (Precision)
for - False omission rate
lrp - Positive likelihood ratio (LR+)
fdr - False discovery rate
npv - Negative predictive value
acc - Accuracy
lrm - Negative likelihood ratio (LR-)
dor - Diagnostic odds ratio
```

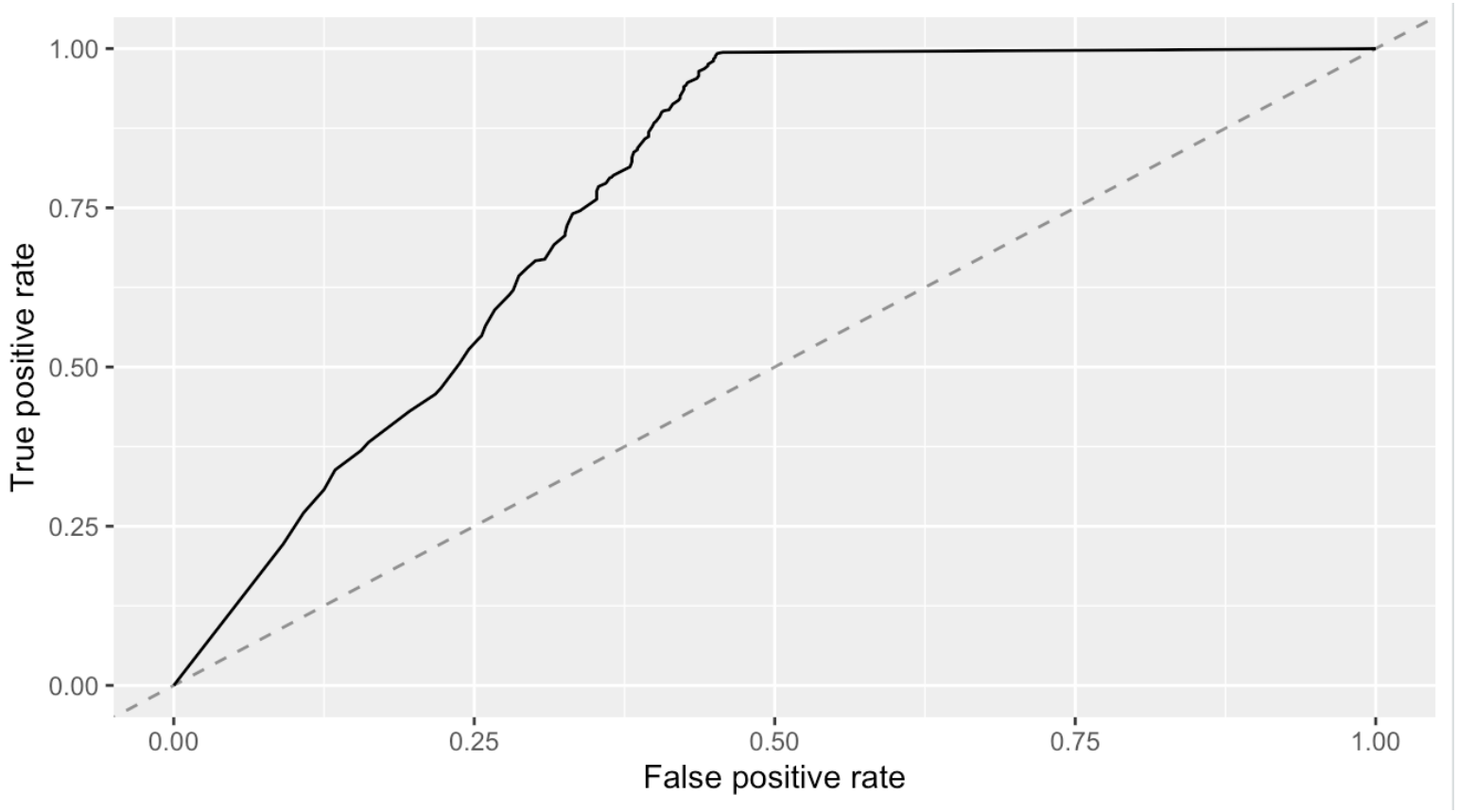
KNN

Confusion Matrix

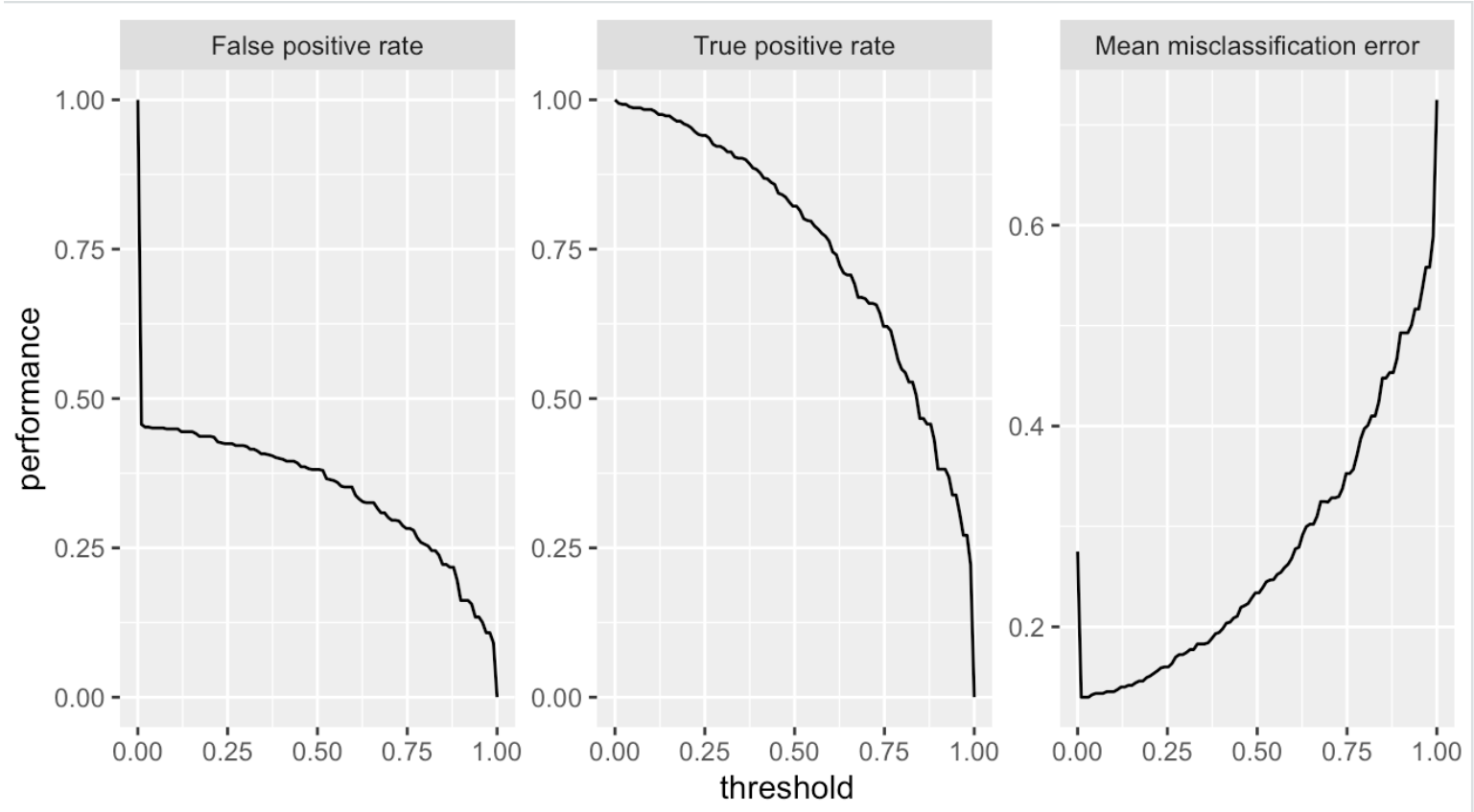
The KNN model has got accuracy of 77% which low compared to Logistic Regression model. The miss rate is little higher comparatively

```
> knn.model1 = train(knn.learner, trainTaskBal1)
> knn.predict1 = predict(knn.model1,testTask1)
> knn.cm1 = calculateROCMeasures(knn.predict1)
> knn.cm1
    predicted
true 1         2
   1 1410      298       tpr: 0.83 fnr: 0.17
   2 240       408       fpr: 0.37 tnr: 0.63
     ppv: 0.85 for: 0.42 lrp: 2.23 acc: 0.77
     fdr: 0.15 npv: 0.58 lrm: 0.28 dor: 8.04


Abbreviations:
tpr - True positive rate (Sensitivity, Recall)
fpr - False positive rate (Fall-out)
fnr - False negative rate (Miss rate)
tnr - True negative rate (Specificity)
ppv - Positive predictive value (Precision)
for - False omission rate
lrp - Positive likelihood ratio (LR+)
fdr - False discovery rate
npv - Negative predictive value
acc - Accuracy
lrm - Negative likelihood ratio (LR-)
dor - Diagnostic odds ratio
```

KNN
ROC Curve



KNN
Performance Chart



Modelling - Random Forest Model

Metrics have come out well and the model is good for this data.

```
> rf.learner = makeLearner("classif.randomForest", predict.type = "prob")
> rf.model= train(rf.learner, trainTaskBal1)
> rf.pred = predict(rf.model, testTask1)
> rf.cm = calculateROCMeasures(rf.pred)
> rf.cm
    predicted
true 1         2
   1 1693      15        tpr: 0.99 fnr: 0.01
   2 293       355       fpr: 0.45 tnr: 0.55
     ppv: 0.85 for: 0.04 lrp: 2.19 acc: 0.87
     fdr: 0.15 npv: 0.96 lrm: 0.02 dor: 136.75


Abbreviations:
tpr - True positive rate (Sensitivity, Recall)
fpr - False positive rate (Fall-out)
fnr - False negative rate (Miss rate)
tnr - True negative rate (Specificity)
ppv - Positive predictive value (Precision)
for - False omission rate
lrp - Positive likelihood ratio (LR+)
fdr - False discovery rate
npv - Negative predictive value
acc - Accuracy
lrm - Negative likelihood ratio (LR-)
dor - Diagnostic odds ratio
```

Modelling - CART model

Confusion Matrix

CART model has higher Accuracy of 87% which very similar to Logistics Regression Model but have less Miss Rate compared to Random Forest Model

```
> cart.learner = makeLearner("classif.rpart",predict.type = "prob")
> cart.model = train(cart.learner, trainTaskBal1)
> cart.predict = predict(cart.model, testTask1)
> cart.cm = calculateROCMeasures(cart.predict)
> cart.cm
    predicted
true 1         2
   1 1708      0      tpr: 1    fnr: 0
   2 298       350    fpr: 0.46 tnr: 0.54
     ppv: 0.85 for: 0 lrp: 2.17 acc: 0.87
     fdr: 0.15 npv: 1 lrm: 0    dor: Inf


Abbreviations:
tpr - True positive rate (Sensitivity, Recall)
fpr - False positive rate (Fall-out)
fnr - False negative rate (Miss rate)
tnr - True negative rate (Specificity)
ppv - Positive predictive value (Precision)
for - False omission rate
lrp - Positive likelihood ratio (LR+)
fdr - False discovery rate
npv - Negative predictive value
acc - Accuracy
lrm - Negative likelihood ratio (LR-)
dor - Diagnostic odds ratio
```

Modelling – SVM

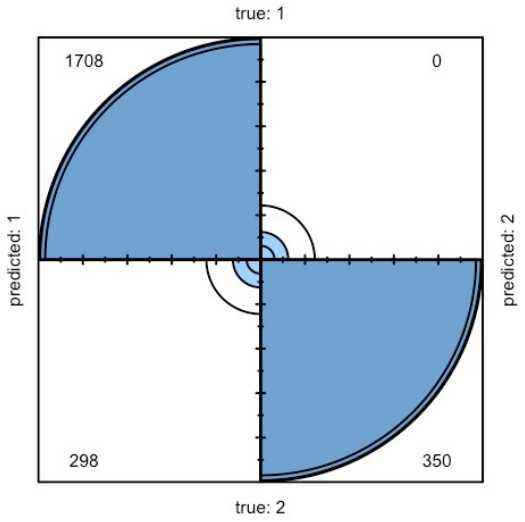
Confusion Matrix

Model has good accuracy of 85% better than KNN model

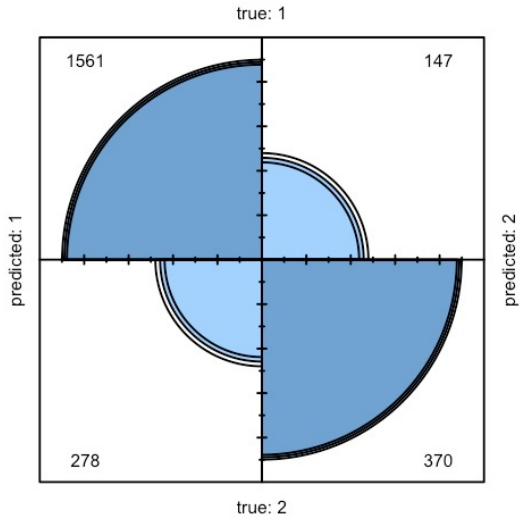
```
> # SVM
> set.seed(124)
> svm.learner = makeLearner("classif.ksvm", predict.type = "prob")
> svm.model= train(svm.learner, trainTaskBal1)
> svm.pred = predict(svm.model, testTask1)
> svm.cm = calculateROCMeasures(svm.pred)
> svm.cm
    predicted
true 1         2
   1 1627      81        tpr: 0.95 fnr: 0.05
   2 282       366       fpr: 0.44 tnr: 0.56
     ppv: 0.85 for: 0.18 lrp: 2.19 acc: 0.85
     fdr: 0.15 npv: 0.82 lrm: 0.08 dor: 26.07


Abbreviations:
tpr - True positive rate (Sensitivity, Recall)
fpr - False positive rate (Fall-out)
fnr - False negative rate (Miss rate)
tnr - True negative rate (Specificity)
ppv - Positive predictive value (Precision)
for - False omission rate
lrp - Positive likelihood ratio (LR+)
fdr - False discovery rate
npv - Negative predictive value
acc - Accuracy
lrm - Negative likelihood ratio (LR-)
dor - Diagnostic odds ratio
```

Comparison of Confusion Matrix

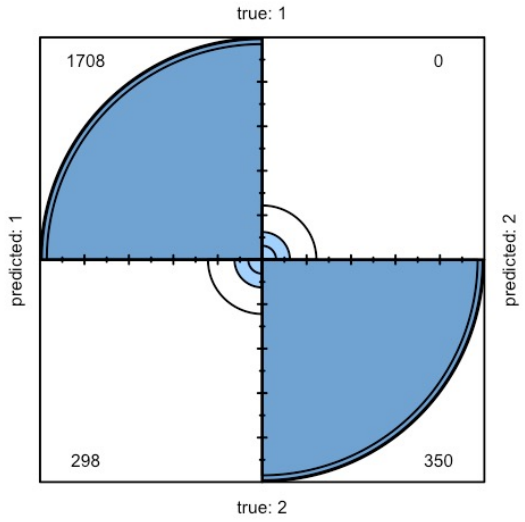
Models	LR	NB	KNN	RF	CART	SVM
Precision	0. 87	0. 85	0. 85	0. 85	0. 85	0. 85
Recall	1. 00	0. 91	0. 83	0. 99	1. 00	0. 95
F1 score	0. 93	0. 88	0. 84	0. 91	0. 92	0. 90
Accuracy	0. 87	0. 82	0. 87	0. 87	0. 85	0. 77



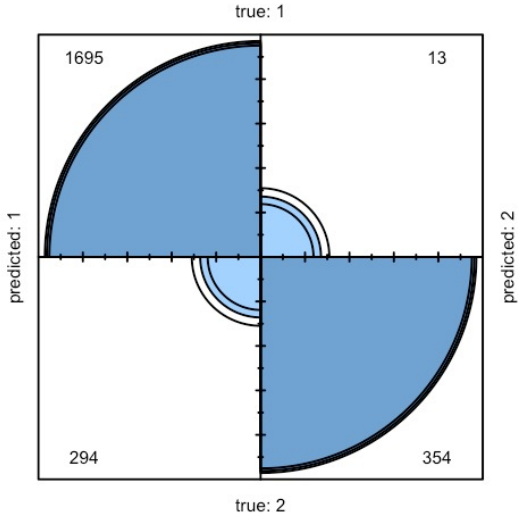
Logistic Regression



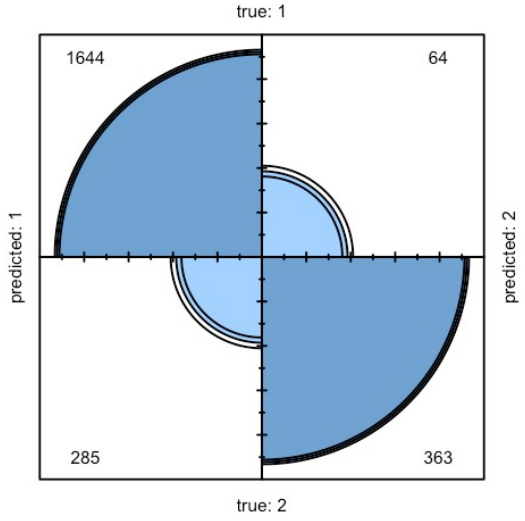
Naïve Bayes



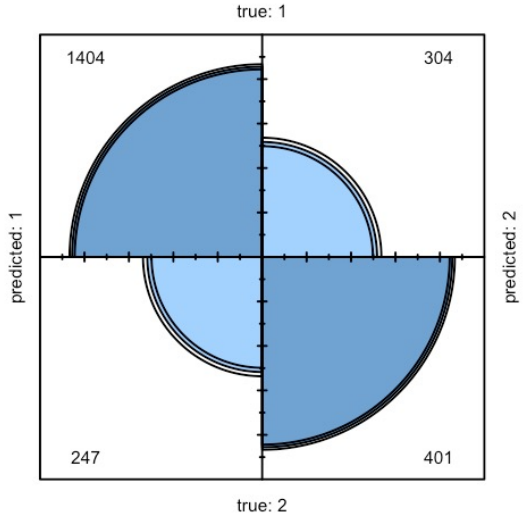
CART



Random Forest



SVM



KNN

Conclusion

We have built several models to predict the shipment modes and below are the metric evaluations:

- With the above result, we should look at F1 score of the models rather than checking the accuracy as the data was heavily imbalanced.
- If we go by F1 score, we see that Logistic Regression model and CART model have done well.

We have built various models to understand the factors which influence the choice of Shipping mode. Using models like RF and CART we found out that most important factors are:

- Sales
- Order Quantity
- Product Container Type – Jumbo Box
- Product Container Type – Jumbo Drum

The above four key factors play an influential role to predict whether the shipment mode is delivery truck or regular air. This is the key takeaway from our analysis.