MINI PROJECT

venkat

13 November 2018

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```r
#LOading data
salesData <- read.csv("C:/Users/venka/OneDrive/Desktop/MINI PROJECT/salesData.csv")


#install packages and load packages
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

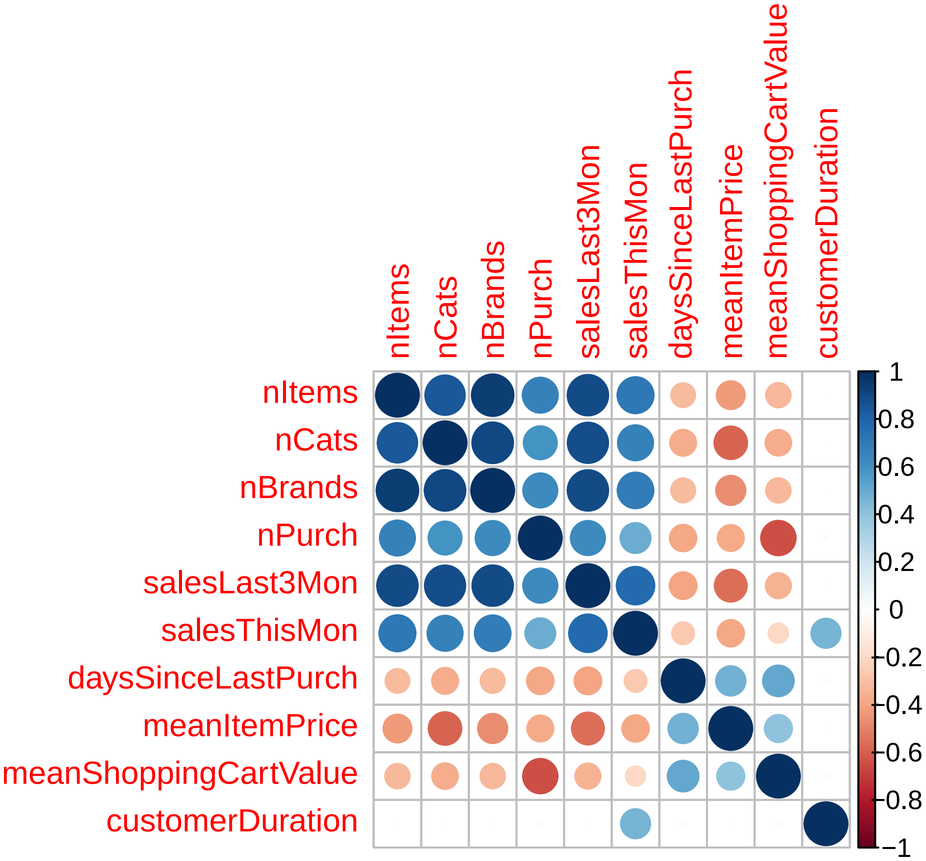
```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

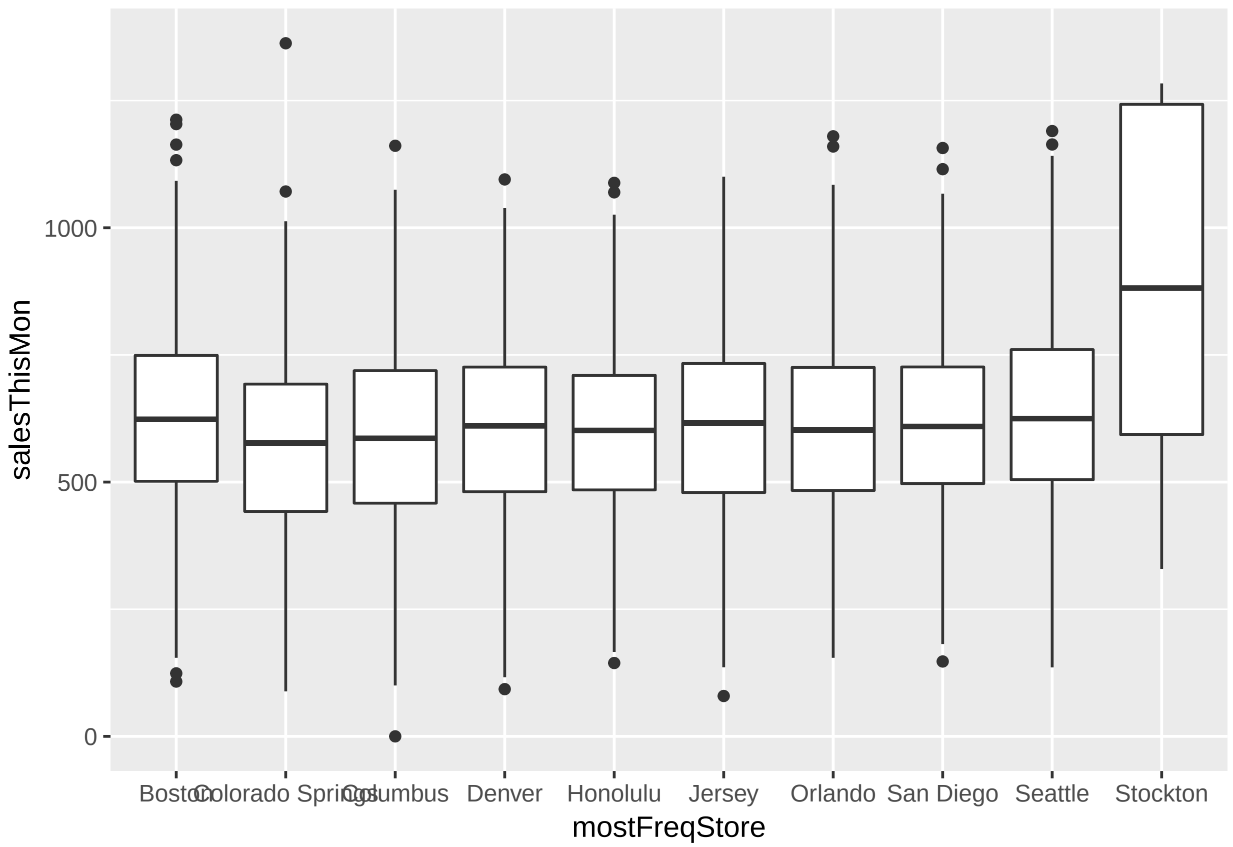
```r
library(ggplot2)

# Structure of dataset
str(salesData, give.attr = FALSE)
```

```
## 'data.frame':    5122 obs. of  14 variables:
##  $ id                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ nItems              : int  1469 1463 262 293 108 216 174 122 204 308 ...
##  $ mostFreqStore       : Factor w/ 10 levels "Boston","Colorado Springs",..: 10 10 2 2 2 1 3 9 6 9
##  $ mostFreqCat         : Factor w/ 10 levels "Alcohol","Baby",..: 1 1 10 3 4 1 8 10 3 1 ...
##  $ nCats               : int  72 73 55 50 32 41 36 31 41 52 ...
##  $ preferredBrand      : Factor w/ 10 levels "Akar","Alekto",..: 10 10 3 10 3 3 3 3 3 3 ...
##  $ nBrands             : int  517 482 126 108 79 98 78 62 99 103 ...
##  $ nPurch              : int  82 88 56 43 18 35 34 12 26 33 ...
##  $ salesLast3Mon       : num  2742 2791 1530 1766 1180 ...
##  $ salesThisMon        : num  1284 1243 683 730 553 ...
##  $ daysSinceLastPurch  : int  1 1 1 1 12 2 2 4 14 1 ...
##  $ meanItemPrice       : num  1.87 1.91 5.84 6.03 10.93 ...
##  $ meanShoppingCartValue: num  33.4 31.7 27.3 41.1 65.6 ...
##  $ customerDuration     : int  821 657 548 596 603 673 612 517 709 480 ...
```

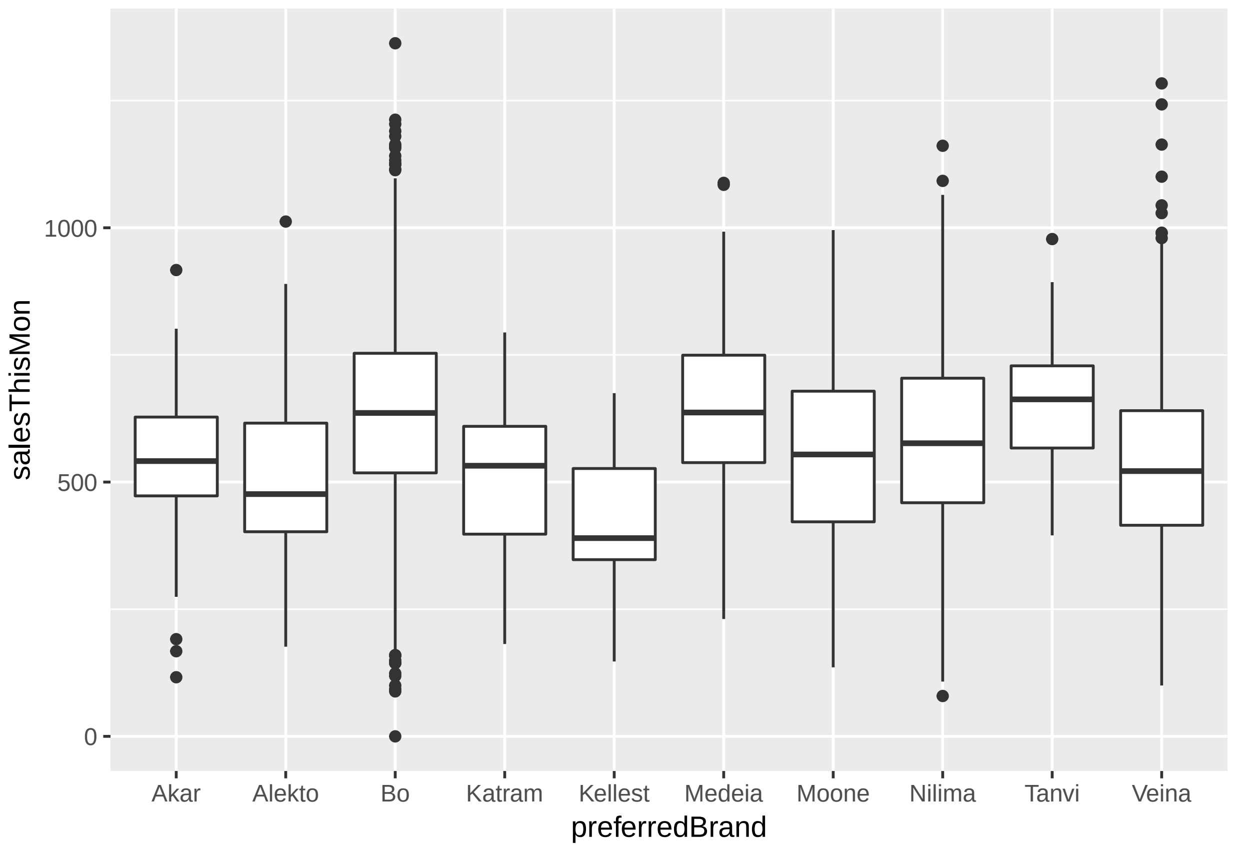
```
View(salesData)
# Visualization of correlations
salesData %>% select_if(is.numeric) %>%
  select(-id) %>%
  cor() %>% corrplot()
```



```
# Frequent stores
ggplot(salesData) +
  geom_boxplot(aes(x = mostFreqStore, y = salesThisMon))
```



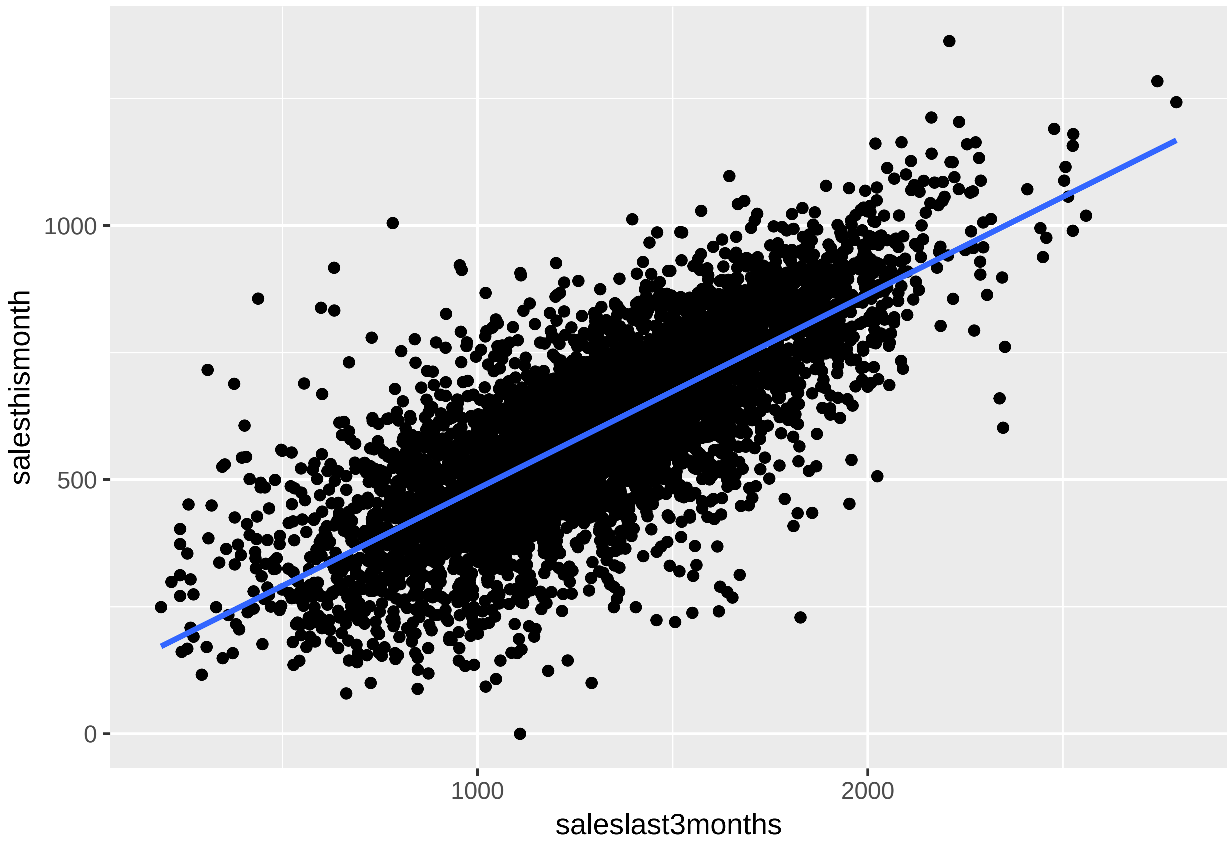
```
# Preferred brand
ggplot(salesData) +
  geom_boxplot(aes(x = preferredBrand, y = salesThisMon))
```



```r
# Model specification using lm
salesSimpleModel <- lm(salesThisMon ~ salesLast3Mon,
                       data = salesData)
# Looking at model summary
summary(salesSimpleModel)
```

```
##
## Call:
## lm(formula = salesThisMon ~ salesLast3Mon, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -570.18  -68.26    3.21   72.98  605.58
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   99.690501   6.083886   16.39   <2e-16 ***
## salesLast3Mon  0.382696   0.004429   86.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 117.5 on 5120 degrees of freedom
## Multiple R-squared:  0.5932, Adjusted R-squared:  0.5931
## F-statistic:  7465 on 1 and 5120 DF,  p-value: < 2.2e-16
```

```r
ggplot(salesData,aes(salesLast3Mon,salesThisMon))+
  geom_point()+
  geom_smooth(method=lm,se=FALSE)+
  xlab("saleslast3months")+
  ylab("salesthismonth")
```



Including Plots

You can also embed plots, for example:

```
#Multiple linear regression

MultipleLM <- lm(salesThisMon ~ salesLast3Mon+id+nItems+mostFreqStore+
                   mostFreqCat+nCats+preferredBrand+nBrands+nPurch+salesLast3Mon+
                   salesThisMon+daysSinceLastPurch+meanItemPrice+meanShoppingCartValue+
                   customerDuration,data=salesData)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 10 in
## model.matrix: no columns are assigned
```

```
summary(MultipleLM)
```

```
##
## Call:
## lm(formula = salesThisMon ~ salesLast3Mon + id + nItems + mostFreqStore +
##     mostFreqCat + nCats + preferredBrand + nBrands + nPurch +
##     salesLast3Mon + salesThisMon + daysSinceLastPurch + meanItemPrice +
##     meanShoppingCartValue + customerDuration, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -322.95  -50.81    0.74   50.87  398.94
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2.614e+02  1.775e+01 -14.726  < 2e-16 ***
## salesLast3Mon                  3.751e-01  8.600e-03  43.614  < 2e-16 ***
## id                             9.248e-04  6.918e-04   1.337 0.181367
## nItems                         1.620e-01  2.711e-02   5.975 2.46e-09 ***
## mostFreqStoreColorado Springs -7.023e+00  4.351e+00  -1.614 0.106591
## mostFreqStoreColumbus          1.194e+00  3.684e+00   0.324 0.745900
## mostFreqStoreDenver           -8.219e+00  5.138e+00  -1.600 0.109710
## mostFreqStoreHonolulu         -1.566e+01  4.919e+00  -3.184 0.001463 **
## mostFreqStoreJersey           -2.158e+01  5.031e+00  -4.291 1.82e-05 ***
## mostFreqStoreOrlando          -1.028e+01  4.496e+00  -2.286 0.022322 *
## mostFreqStoreSan Diego        -1.989e+01  5.718e+00  -3.479 0.000507 ***
## mostFreqStoreSeattle          -9.585e+00  3.541e+00  -2.707 0.006822 **
## mostFreqStoreStockton         -1.156e+02  3.583e+01  -3.225 0.001268 **
## mostFreqCatBaby               -3.453e+00  3.513e+00  -0.983 0.325620
## mostFreqCatBakery             -1.025e+01  5.456e+00  -1.878 0.060376 .
## mostFreqCatBeverages           3.728e-01  7.007e+00   0.053 0.957574
## mostFreqCatClothes            -8.677e+00  6.214e+00  -1.396 0.162667
## mostFreqCatFresh food         -6.299e+00  7.244e+00  -0.869 0.384642
## mostFreqCatFrozen food        -8.083e+00  3.840e+00  -2.105 0.035322 *
## mostFreqCatPackaged food      -9.868e-01  4.357e+00  -0.226 0.820838
## mostFreqCatPets                8.664e+00  7.242e+00   1.196 0.231633
## mostFreqCatShoes               3.327e+00  3.285e+00   1.013 0.311294
## nCats                         -7.828e-01  2.346e-01  -3.336 0.000855 ***
## preferredBrandAlekto          -5.085e+00  1.649e+01  -0.308 0.757863
## preferredBrandBo              -2.466e+01  1.438e+01  -1.715 0.086432 .
## preferredBrandKatram          -6.272e+01  2.333e+01  -2.688 0.007213 **
## preferredBrandKellest         -5.288e+01  2.214e+01  -2.388 0.016955 *
## preferredBrandMedeia          -2.116e+01  1.556e+01  -1.360 0.173856
## preferredBrandMoone           -4.103e+01  1.627e+01  -2.522 0.011711 *
## preferredBrandNilima          -2.843e+01  1.454e+01  -1.955 0.050631 .
## preferredBrandTanvi            3.260e+01  2.131e+01   1.530 0.126133
## preferredBrandVeina           -1.818e+01  1.452e+01  -1.252 0.210471
## nBrands                       -5.314e-02  8.476e-02  -0.627 0.530745
## nPurch                         4.767e-01  1.513e-01   3.151 0.001636 **
## daysSinceLastPurch             1.801e-01  1.524e-01   1.181 0.237512
## meanItemPrice                  1.779e-01  9.289e-02   1.915 0.055532 .
## meanShoppingCartValue          2.593e-01  2.618e-02   9.905  < 2e-16 ***
## customerDuration               5.713e-01  7.147e-03  79.938  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.34 on 5084 degrees of freedom
## Multiple R-squared:  0.825,  Adjusted R-squared:  0.8237
## F-statistic: 647.7 on 37 and 5084 DF,  p-value: < 2.2e-16
```

```r
library(rms)
```

```
## Loading required package: Hmisc
```

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

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
## Loading required package: SparseM
```

```
##
## Attaching package: 'SparseM'
```

```
## The following object is masked from 'package:base':
##
##     backsolve
```

```r
vif(MultipleLM)
```

```
##                 salesLast3Mon                            id
##                      8.702133                      1.014808
##                        nItems mostFreqStoreColorado Springs
##                     11.793761                      1.479010
##         mostFreqStoreColumbus           mostFreqStoreDenver
##                      1.750119                      1.293184
##         mostFreqStoreHonolulu           mostFreqStoreJersey
##                      1.339893                      1.317457
##          mostFreqStoreOrlando        mostFreqStoreSan Diego
##                      1.403759                      1.220688
##          mostFreqStoreSeattle         mostFreqStoreStockton
##                      1.798055                      1.072300
##               mostFreqCatBaby             mostFreqCatBakery
##                      1.457026                      1.246035
##          mostFreqCatBeverages            mostFreqCatClothes
##                      1.079024                      1.157217
##         mostFreqCatFresh food        mostFreqCatFrozen food
##                      1.070048                      1.296358
##      mostFreqCatPackaged food               mostFreqCatPets
##                      1.268867                      1.077766
##              mostFreqCatShoes                         nCats
##                      1.417865                      8.408871
##          preferredBrandAlekto              preferredBrandBo
##                      3.846192                     41.092962
##          preferredBrandKatram         preferredBrandKellest
##                      1.632990                      1.714237
##          preferredBrandMedeia           preferredBrandMoone
##                      6.123156                      4.595465
##          preferredBrandNilima           preferredBrandTanvi
##                     22.726586                      1.889467
##           preferredBrandVeina                       nBrands
##                     20.749246                     14.179569
##                        nPurch            daysSinceLastPurch
##                      3.084009                      1.585075
##                 meanItemPrice         meanShoppingCartValue
##                      1.987908                      2.247795
##              customerDuration
##                      1.004680
```

```r
# Estimating the full model
salesModel1 <- lm(salesThisMon ~ . - id,
                  data = salesData)

# Checking variance inflation factors
vif(salesModel1)
```

```
##                        nItems mostFreqStoreColorado Springs
##                     11.772600                      1.478098
##         mostFreqStoreColumbus           mostFreqStoreDenver
##                      1.746101                      1.289203
##         mostFreqStoreHonolulu           mostFreqStoreJersey
##                      1.338330                      1.317158
##          mostFreqStoreOrlando        mostFreqStoreSan Diego
##                      1.401396                      1.219922
##          mostFreqStoreSeattle         mostFreqStoreStockton
##                      1.794891                      1.070250
##               mostFreqCatBaby             mostFreqCatBakery
##                      1.456921                      1.246035
##          mostFreqCatBeverages            mostFreqCatClothes
##                      1.079007                      1.156841
##         mostFreqCatFresh food        mostFreqCatFrozen food
##                      1.069987                      1.296358
##      mostFreqCatPackaged food               mostFreqCatPets
##                      1.268000                      1.077488
##              mostFreqCatShoes                         nCats
##                      1.417807                      8.402073
##          preferredBrandAlekto              preferredBrandBo
##                      3.844176                     41.075930
##          preferredBrandKatram         preferredBrandKellest
##                      1.632978                      1.713510
##          preferredBrandMedeia           preferredBrandMoone
##                      6.120384                      4.591570
##          preferredBrandNilima           preferredBrandTanvi
##                     22.714376                      1.885777
##           preferredBrandVeina                       nBrands
##                     20.739114                     14.150868
##                        nPurch                 salesLast3Mon
##                      3.083952                      8.697663
##            daysSinceLastPurch                 meanItemPrice
##                      1.585057                      1.987665
##         meanShoppingCartValue              customerDuration
##                      2.247579                      1.004664
```

```
# Estimating new model by removing information on brand
salesModel2 <- lm(salesThisMon ~ . - id-preferredBrand-nBrands,
                  data = salesData)

# Checking variance inflation factors

vif(salesModel2)
```

```
##                        nItems mostFreqStoreColorado Springs
##                      6.987456                      1.470508
##         mostFreqStoreColumbus           mostFreqStoreDenver
##                      1.737790                      1.283222
##         mostFreqStoreHonolulu           mostFreqStoreJersey
##                      1.335457                      1.299889
##          mostFreqStoreOrlando        mostFreqStoreSan Diego
##                      1.398318                      1.213865
##          mostFreqStoreSeattle         mostFreqStoreStockton
##                      1.788777                      1.052065
##               mostFreqCatBaby             mostFreqCatBakery
##                      1.412755                      1.236939
##          mostFreqCatBeverages            mostFreqCatClothes
##                      1.077907                      1.105054
##         mostFreqCatFresh food        mostFreqCatFrozen food
##                      1.067089                      1.270953
##      mostFreqCatPackaged food               mostFreqCatPets
##                      1.235165                      1.072278
##              mostFreqCatShoes                         nCats
##                      1.384861                      5.813494
##                        nPurch                 salesLast3Mon
##                      3.069046                      8.412520
##            daysSinceLastPurch                 meanItemPrice
##                      1.579426                      1.925494
##         meanShoppingCartValue              customerDuration
##                      2.238410                      1.002981
```

```
# getting an overview of new data
salesData2_4 <- read.csv("C:/Users/venka/OneDrive/Desktop/MINI PROJECT/salesData2_4.csv")
head(salesData2_4)
```

```
##   id nItems    mostFreqStore   mostFreqCat nCats preferredBrand nBrands
## 1  1   1401         Stockton       Alcohol    73          Veina     483
## 2  2   1461         Stockton       Alcohol    74          Veina     484
## 3  3    262 Colorado Springs         Shoes    55             Bo     131
## 4  4    250 Colorado Springs        Bakery    43          Veina      93
## 5  5    149 Colorado Springs Packaged food    36             Bo      90
## 6  6    208           Boston         Shoes    35             Bo      82
##   nPurch salesLast3Mon daysSinceLastPurch meanItemPrice
## 1     85       2712.99                  3      1.936467
## 2     86       2744.57                  2      1.878556
## 3     55       1527.10                  1      5.828626
## 4     44       1675.11                  2      6.700440
## 5     27       1265.18                  4      8.491141
## 6     33       1353.23                  1      6.505913
##   meanShoppingCartValue customerDuration
## 1              31.91753              852
## 2              31.91360              688
## 3              27.76545              579
## 4              38.07068              627
## 5              46.85852              634
## 6              41.00697              704
```

```
summary(salesData2_4)
```

```
##        id           nItems                mostFreqStore
##  Min.   :   1   Min.   :   1.0   Seattle         :1104
##  1st Qu.:1372   1st Qu.:  84.0   Columbus        : 952
##  Median :2733   Median : 155.0   Boston          : 873
##  Mean   :2729   Mean   : 185.9   Colorado Springs: 530
##  3rd Qu.:4085   3rd Qu.: 257.0   Orlando         : 467
##  Max.   :5455   Max.   :1461.0   Honolulu        : 359
##                                  (Other)         : 888
##         mostFreqCat       nCats       preferredBrand    nBrands
##  Alcohol      :1506   Min.   : 1.00   Bo     :3328   Min.   :  1.00
##  Shoes        : 930   1st Qu.:27.00   Nilima : 771   1st Qu.: 45.00
##  Baby         : 857   Median :37.00   Veina  : 709   Median : 75.00
##  Frozen food  : 549   Mean   :36.23   Medeia : 152   Mean   : 81.66
##  Packaged food: 471   3rd Qu.:46.00   Moone  :  75   3rd Qu.:110.00
##  Bakery       : 276   Max.   :74.00   Alekto :  61   Max.   :484.00
##  (Other)      : 584                   (Other):  77
##      nPurch      salesLast3Mon  daysSinceLastPurch meanItemPrice
##  Min.   : 1.00   Min.   : 189   Min.   : 1.000     Min.   :  1.879
##  1st Qu.:11.00   1st Qu.:1068   1st Qu.: 2.000     1st Qu.:  6.049
##  Median :17.00   Median :1331   Median : 4.000     Median :  8.556
##  Mean   :20.02   Mean   :1324   Mean   : 6.589     Mean   : 12.116
##  3rd Qu.:27.00   3rd Qu.:1570   3rd Qu.: 7.000     3rd Qu.: 12.969
##  Max.   :86.00   Max.   :2745   Max.   :87.000     Max.   :313.050
##
##  meanShoppingCartValue customerDuration
##  Min.   :  17.58       Min.   :  31.0
##  1st Qu.:  53.88       1st Qu.: 580.0
##  Median :  75.77       Median : 682.0
##  Mean   :  91.88       Mean   : 676.8
##  3rd Qu.: 109.74       3rd Qu.: 777.0
##  Max.   :1147.66       Max.   :1386.0
##
```

```
# predicting sales
predSales5 <- predict(salesModel2, newdata = salesData2_4)

# calculating mean of future sales
mean(predSales5 ,na.rm=TRUE)
```

```
## [1] 625.1438
```