Sales Predictor

Christopher Himmel

January 16, 2016

# Introduction

A valuable exercise in retail is being able to predict sales of your inventory. Predictive models can be used to do this, building a relationship betwen given information that you have currently and future sales. Current information includes previous sales and Item attributes, of which there are many.

The many item attributes are recorded when the new SKU's are entered by hand one by one, or loaded as a group by uploading them and processing them into the system. They are stored in the Inventory Master file (INVMST). Sales history is loaded as stores sell inventory, quantities and prices are uploaded from the Point-Of-Sale and accumulated into a few buckets in the Inventory Balance files (INVBAL and INVCBL).

The data used for this exercise is directly pulled out of JDA's Merchandise Management System, which runs on IBM's System-i. The information is downloaded from INVMST, INVCBL and INVBAL (limited to only ISTORE between 1001 and 1003).

The following is the R code and results from this exercise.

# Set up R environment

#Needed<-c("dplyr", "tm", "wordcloud", "SnowballC", "corrplot", "rpart", "rattle", "randomForest", "tidyr", "ggplot2")  
#install.packages(Needed, dependencies=TRUE)  
library(tm)  
library(dplyr)  
library(wordcloud)  
library(SnowballC)  
library(corrplot)  
library(rpart)  
library(rattle)

## Warning: package 'rattle' was built under R version 3.2.5

library(randomForest)  
library(tidyr)  
library(ggplot2)  
setwd("C:/Users/Christopher Himmel/Dropbox/Deep Learning/SlideRule Foundations/Capstone")

# Upload data pulled out of MMS

Load and view Item Master (INVMST):

Item\_Master = read.csv("INVMST - Item Master.csv")  
#View(Item\_Master)  
#summary(Item\_Master)

Load and view Chain Level Inventory Balance file (INVCBL):

Inv\_Bal\_Chain = read.csv("INVCBL - Chain Level Inventory Balance Data.csv")  
#View(Inv\_Bal\_Chain)  
#summary(Inv\_Bal\_Chain)

Load and view Store Level Inventory Balance file, filtered for Niquea'D stores (INVBAL):

Inv\_Bal\_Store = read.csv("INVBAL - Store Level Inventory Balance Data - Niquea'D.csv")  
#View(Inv\_Bal\_Store)  
#summary(Inv\_Bal\_Store)

# Format data set

Create chain level field with combined regular sales and advertised sales:

Inv\_Bal\_Chain\_mut <-  
 Inv\_Bal\_Chain %>%  
 mutate(chn\_sales=CBRSUY+CBASUY)

Create total Niquea'D summarized sales with combined regular sales and advertised sales:

Inv\_Bal\_Store\_sum <-   
 Inv\_Bal\_Store %>%  
 group\_by(INUMBR) %>%  
 summarise(nd\_sales=sum(IBRSUY+IBASUY),  
 IBHAND\_sum=sum(IBHAND),  
 IBWKCR\_sum=sum(IBWKCR),  
 IBWK01\_sum=sum(IBWK01),  
 IBWK02\_sum=sum(IBWK02),  
 IBWK03\_sum=sum(IBWK03),  
 IBWK04\_sum=sum(IBWK04),  
 IBWK05\_sum=sum(IBWK05),  
 IBWK06\_sum=sum(IBWK06),  
 IBWK07\_sum=sum(IBWK07),  
 IBWK08\_sum=sum(IBWK08))  
Inv\_Bal\_Store\_sum <- subset(Inv\_Bal\_Store\_sum, IBHAND\_sum!=0 | nd\_sales!=0)

Combine two new sales numbers into one file by SKU:

Inv\_values <-  
 Inv\_Bal\_Store\_sum %>%  
 left\_join(Inv\_Bal\_Chain\_mut, by="INUMBR") %>%  
 select(INUMBR,nd\_sales,chn\_sales,IBHAND\_sum,IBWKCR\_sum,IBWK01\_sum,IBWK02\_sum,IBWK03\_sum,IBWK04\_sum,  
 IBWK05\_sum,IBWK06\_sum,IBWK07\_sum,IBWK08\_sum)

Combine SKU, sales data into one file for analysis:

Inv\_final <-  
 Inv\_values %>%  
 left\_join(Item\_Master, by="INUMBR")

# Break descriptions down for learning

Extract word list from description:

word\_list <- paste(Inv\_final$IDESCR, collapse=" ")  
word\_list\_vector <- VectorSource(word\_list)  
rm(word\_list)  
word\_list\_corpus <- Corpus(word\_list\_vector)  
rm(word\_list\_vector)  
  
word\_list\_corpus <- tm\_map(word\_list\_corpus, content\_transformer(tolower))  
word\_list\_corpus <- tm\_map(word\_list\_corpus, removePunctuation)  
word\_list\_corpus <- tm\_map(word\_list\_corpus, stripWhitespace)  
word\_list\_corpus <- tm\_map(word\_list\_corpus, removeNumbers)  
word\_list\_corpus <- tm\_map(word\_list\_corpus, removeWords, stopwords("english"))  
word\_list\_corpus <- tm\_map(word\_list\_corpus, stemDocument)  
  
tdm<-TermDocumentMatrix(word\_list\_corpus)  
rm(word\_list\_corpus)  
tdm2<-as.matrix(tdm)  
rm(tdm)  
tdm3<-as.data.frame(tdm2)  
rm(tdm2)  
colnames(tdm3)<-c("count")  
tdm4<-subset(tdm3,tdm3$count>132)  
rm(tdm3)  
top\_words\_list<-rownames(tdm4)  
rm(tdm4)

Generate training vector:

description\_list <- Inv\_final$IDESCR  
description\_list\_vector <- VectorSource(description\_list)  
rm(description\_list)  
description\_list\_corpus <- Corpus(description\_list\_vector)  
rm(description\_list\_vector)  
  
description\_list\_corpus <- tm\_map(description\_list\_corpus, content\_transformer(tolower))  
description\_list\_corpus <- tm\_map(description\_list\_corpus, removePunctuation)  
description\_list\_corpus <- tm\_map(description\_list\_corpus, stripWhitespace)  
description\_list\_corpus <- tm\_map(description\_list\_corpus, removeNumbers)  
description\_list\_corpus <- tm\_map(description\_list\_corpus, removeWords, stopwords("english"))  
description\_list\_corpus <- tm\_map(description\_list\_corpus, stemDocument)  
  
dtm\_dataset<-DocumentTermMatrix(description\_list\_corpus)  
rm(description\_list\_corpus)  
dtm\_dataset2<-as.matrix(dtm\_dataset)  
rm(dtm\_dataset)  
dtm\_match<-match(top\_words\_list,colnames(dtm\_dataset2))  
dtm\_dataset\_top<-dtm\_dataset2[,dtm\_match]  
rm(dtm\_dataset2)  
dtm\_dataset\_topdf<-as.data.frame(dtm\_dataset\_top)  
rm(dtm\_dataset\_top)  
colnames(dtm\_dataset\_topdf)<-paste("w", colnames(dtm\_dataset\_topdf), sep="\_")

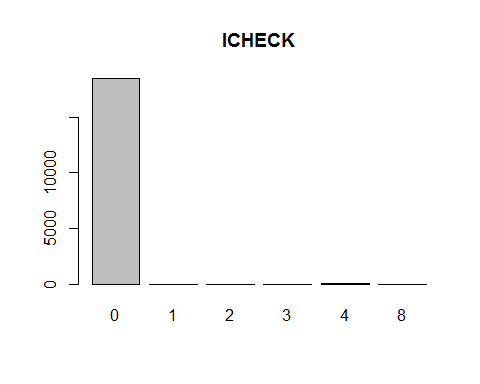
Add word values to training set:

Inv\_final$rownumber<-c(1:nrow(Inv\_final))  
dtm\_dataset\_topdf$rownumber<-rownames(dtm\_dataset\_topdf)  
Inv\_final<-merge(Inv\_final,dtm\_dataset\_topdf,by="rownumber")

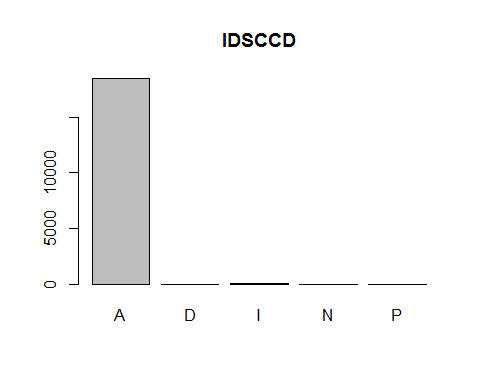
# Test each attribute for significant amount of data

Visualize each column of data:

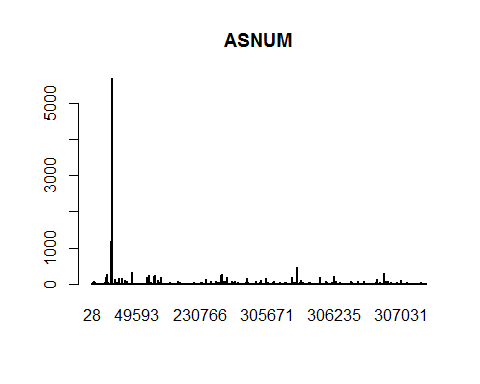
# IDESCR - factors of descriptions  
# ICHECK - all Zeros  
barplot(table(Inv\_final$ICHECK),main="ICHECK")



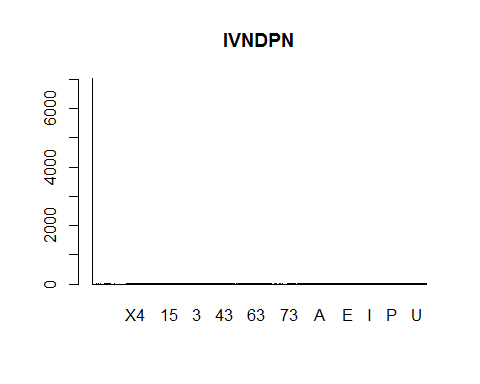
# IDSCCD - minimal values  
barplot(table(Inv\_final$IDSCCD),main="IDSCCD")



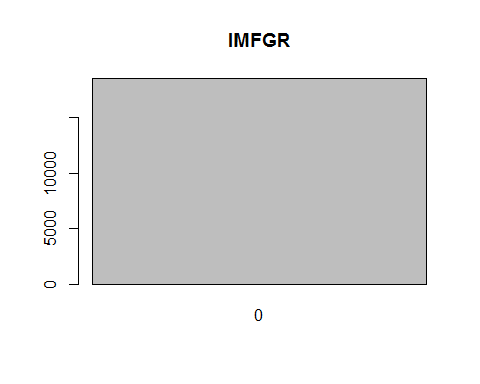
# ISORT - factors of descriptions  
# ISTYLN - all NA's  
# ASNUM - distributed values  
barplot(table(Inv\_final$ASNUM),main="ASNUM")



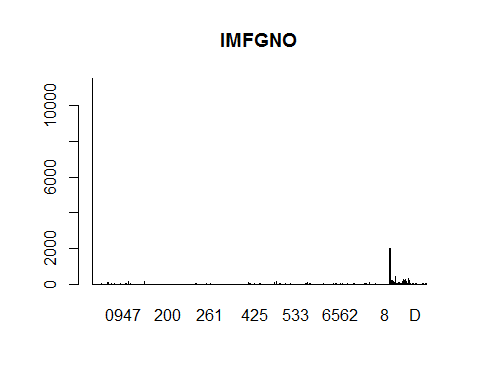
# IVNDPN - empty  
barplot(table(Inv\_final$IVNDPN),main="IVNDPN")



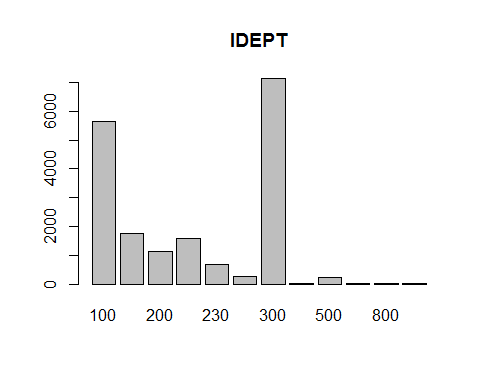
# IMFGR - empty  
barplot(table(Inv\_final$IMFGR),main="IMFGR")



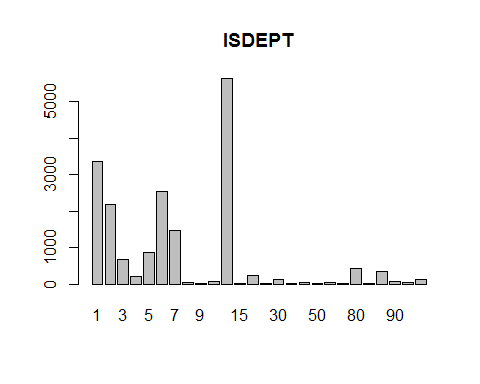
# IMFGNO - some values  
barplot(table(Inv\_final$IMFGNO),main="IMFGNO")



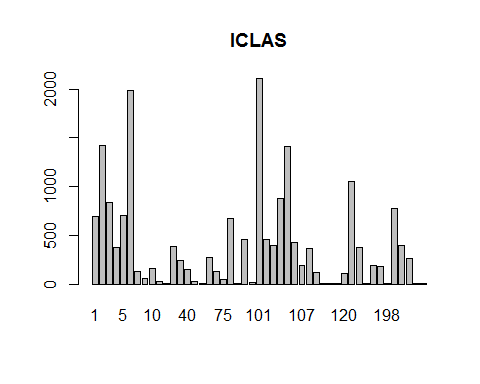
# IDEPT - use  
barplot(table(Inv\_final$IDEPT),main="IDEPT")



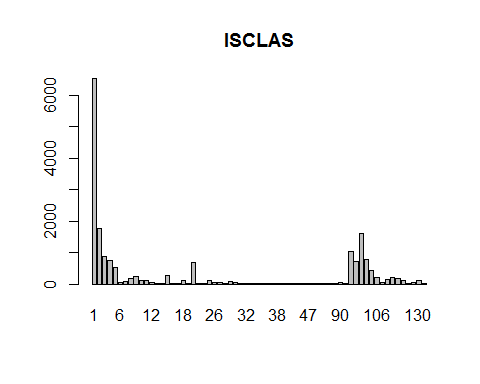
# ISDEPT - use  
barplot(table(Inv\_final$ISDEPT),main="ISDEPT")



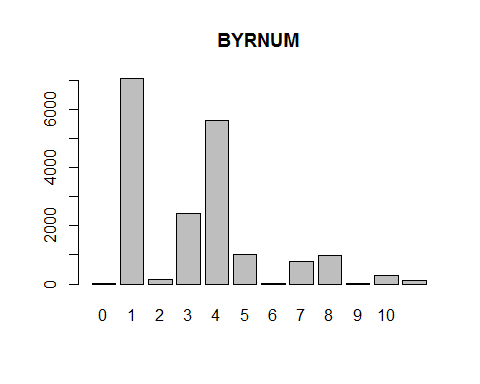
# ICLAS - use  
barplot(table(Inv\_final$ICLAS),main="ICLAS")



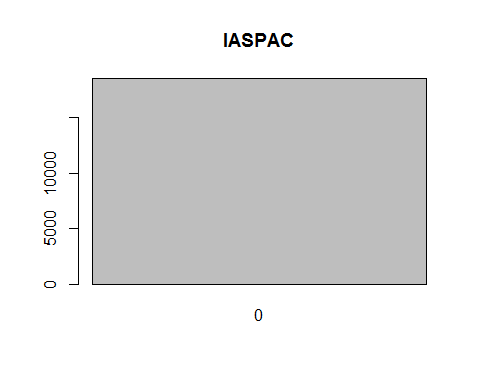
# ISCLAS - use  
barplot(table(Inv\_final$ISCLAS),main="ISCLAS")



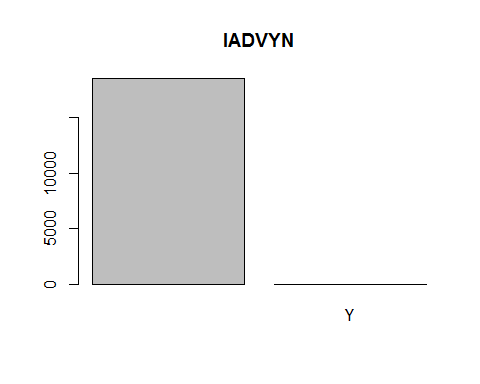
# BYRNUM - convert NA's to 0  
Inv\_final$BYRNUM[is.na(Inv\_final$BYRNUM)]<-0  
barplot(table(Inv\_final$BYRNUM),main="BYRNUM")



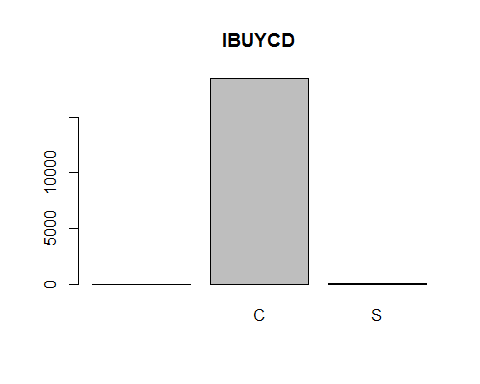
# IASPAC - empty  
barplot(table(Inv\_final$IASPAC),main="IASPAC")



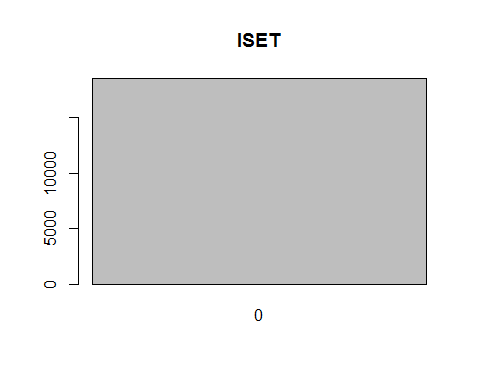
# IADVYN - empty  
barplot(table(Inv\_final$IADVYN),main="IADVYN")



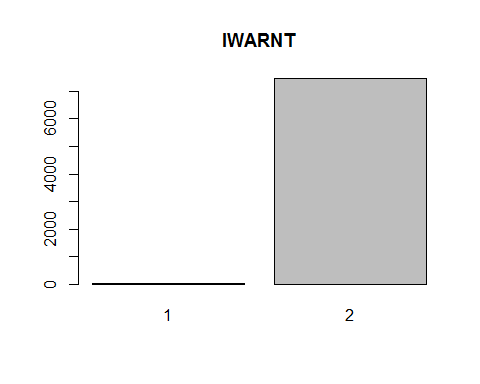
# IBUYCD - mostly empty  
barplot(table(Inv\_final$IBUYCD),main="IBUYCD")



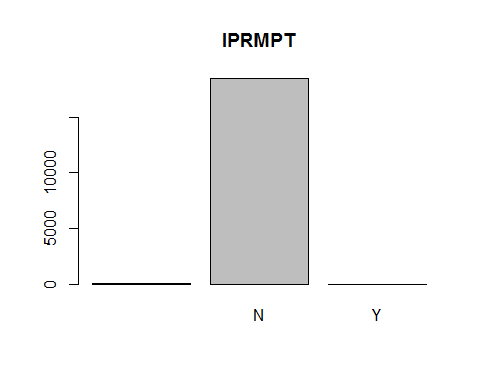
# ISET - empty  
barplot(table(Inv\_final$ISET),main="ISET")



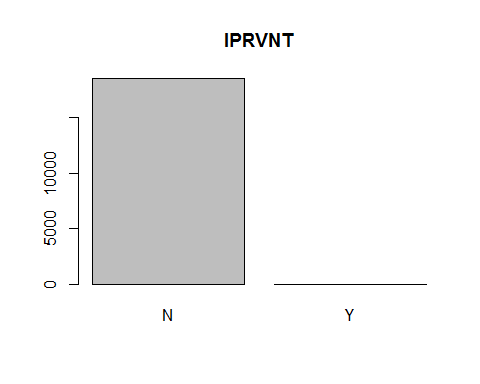
# IWARNT - mostly empty  
barplot(table(Inv\_final$IWARNT),main="IWARNT")



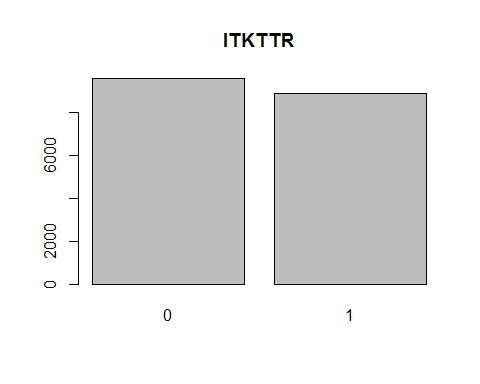
# IPRMPT - mostly empty  
barplot(table(Inv\_final$IPRMPT),main="IPRMPT")



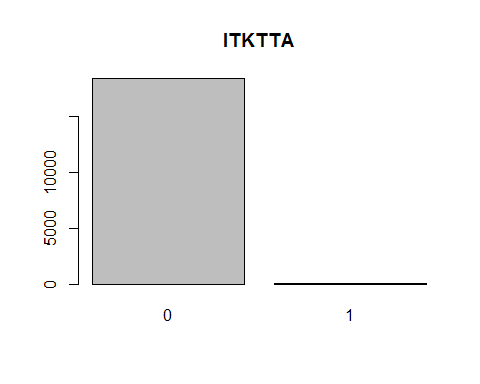
# IPRVNT - mostly empty  
barplot(table(Inv\_final$IPRVNT),main="IPRVNT")



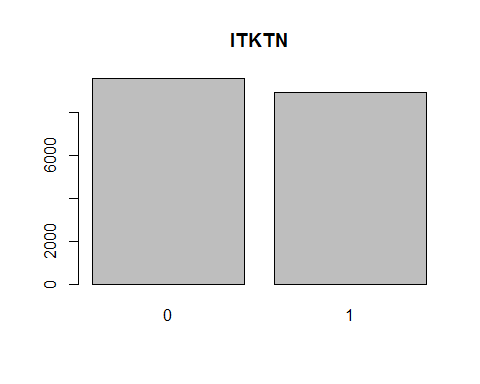
# ITKTTR - convert NA's to 0  
Inv\_final$ITKTTR[is.na(Inv\_final$ITKTTR)]<-0  
barplot(table(Inv\_final$ITKTTR),main="ITKTTR")



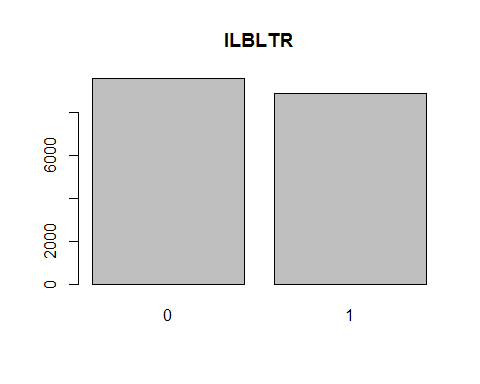
# ITKTTA - mostly empty  
barplot(table(Inv\_final$ITKTTA),main="ITKTTA")



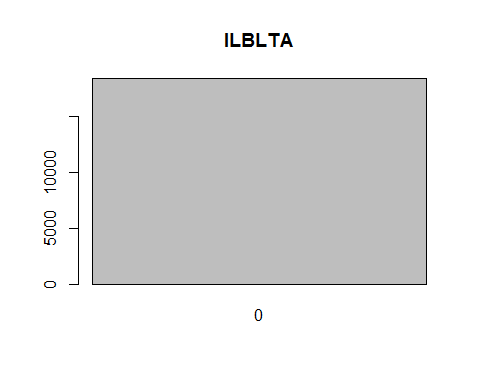
# ITKTN  
barplot(table(Inv\_final$ITKTN),main="ITKTN")



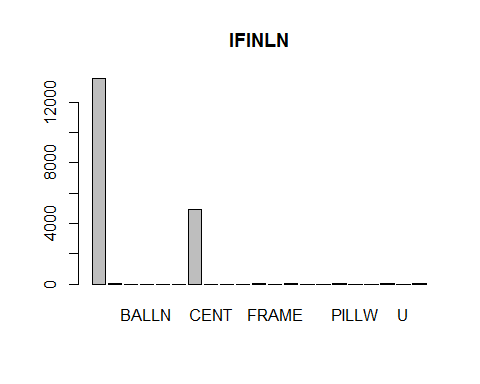
# ILBLTR - convert NA's to 0  
Inv\_final$ILBLTR[is.na(Inv\_final$ILBLTR)]<-0  
barplot(table(Inv\_final$ILBLTR),main="ILBLTR")



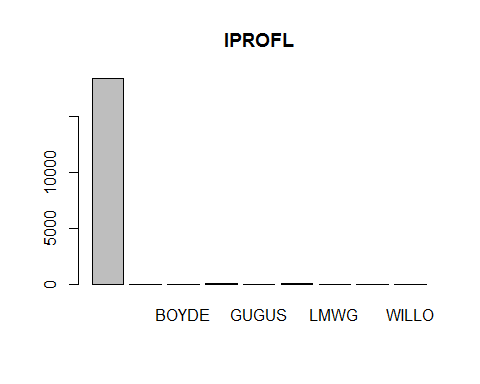
# ILBLTA - empty  
barplot(table(Inv\_final$ILBLTA),main="ILBLTA")



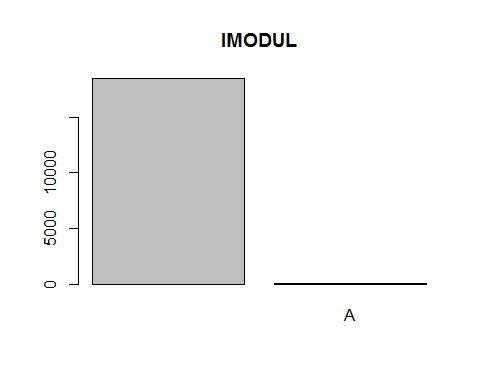
# IFINLN  
barplot(table(Inv\_final$IFINLN),main="IFINLN")



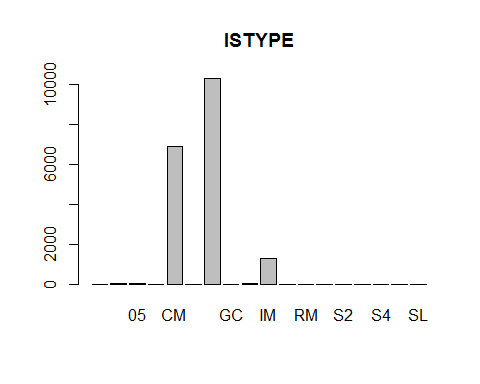
# IPROFL - mostly empty  
barplot(table(Inv\_final$IPROFL),main="IPROFL")



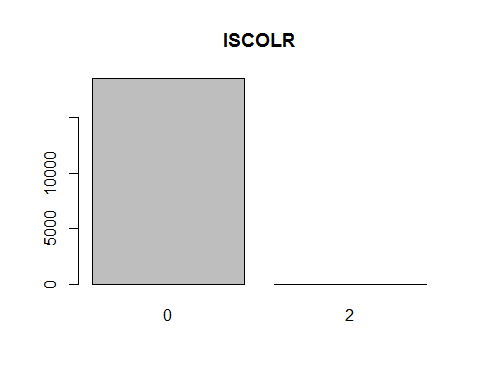
# IMODUL - mostly empty  
barplot(table(Inv\_final$IMODUL),main="IMODUL")



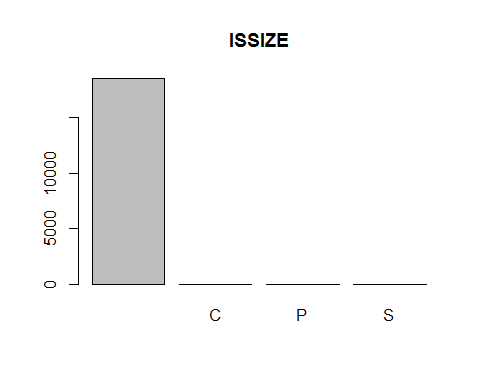
# ISTYPE  
barplot(table(Inv\_final$ISTYPE),main="ISTYPE")



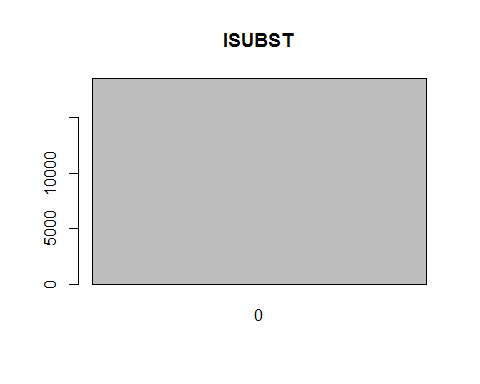
# ISCOLR - mostly empty  
barplot(table(Inv\_final$ISCOLR),main="ISCOLR")



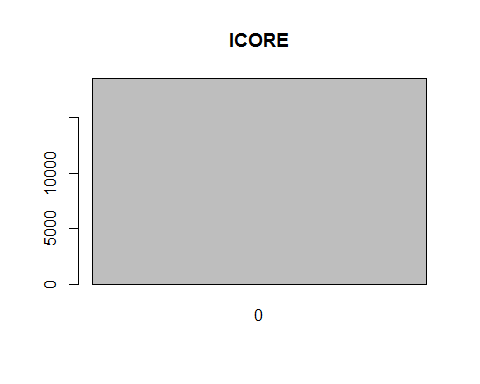
# ISSIZE - mostly empty  
barplot(table(Inv\_final$ISSIZE),main="ISSIZE")



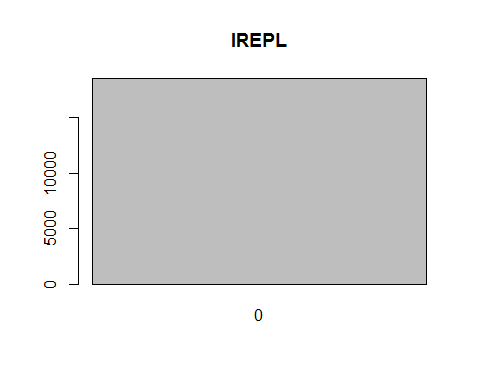
# IHAZCD - all NA's  
# MCHNUM - all NA's  
# ISUBST - empty  
barplot(table(Inv\_final$ISUBST),main="ISUBST")



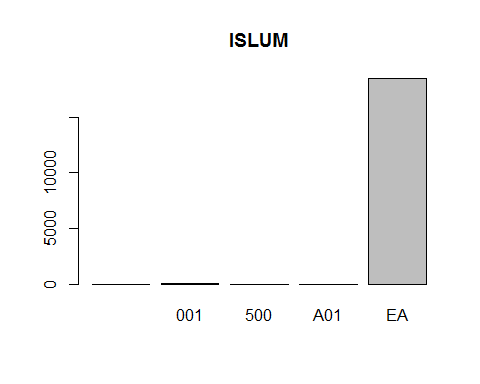
# ICORE - empty  
barplot(table(Inv\_final$ICORE),main="ICORE")



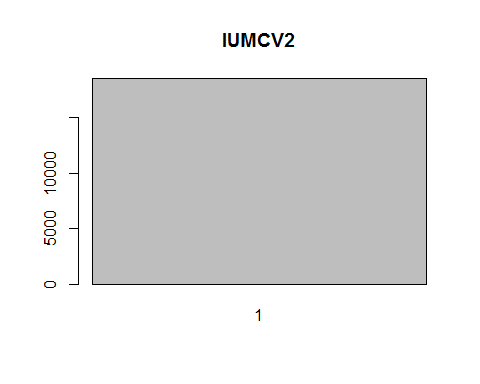
# IREPL - empty  
barplot(table(Inv\_final$IREPL),main="IREPL")



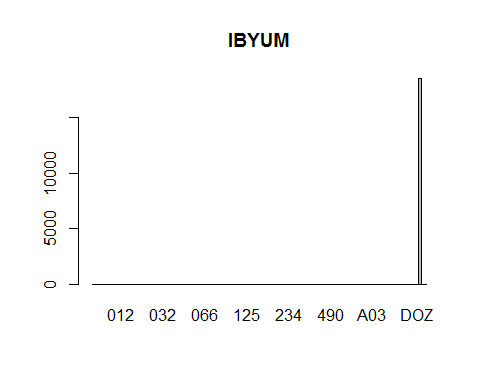
# ISLUM - mostly empty  
barplot(table(Inv\_final$ISLUM),main="ISLUM")



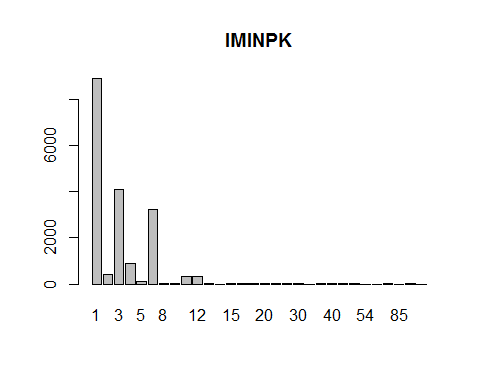
# IUMCV2 - all 1's  
barplot(table(Inv\_final$IUMCV2),main="IUMCV2")



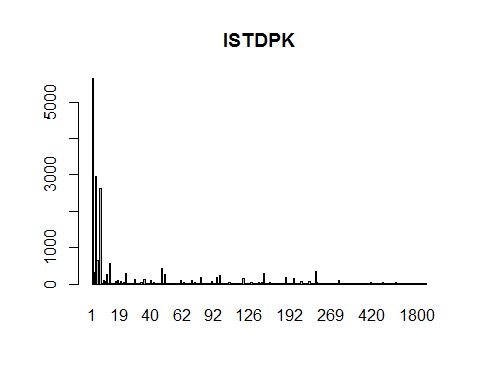
# IBYUM - all EA's  
barplot(table(Inv\_final$IBYUM),main="IBYUM")



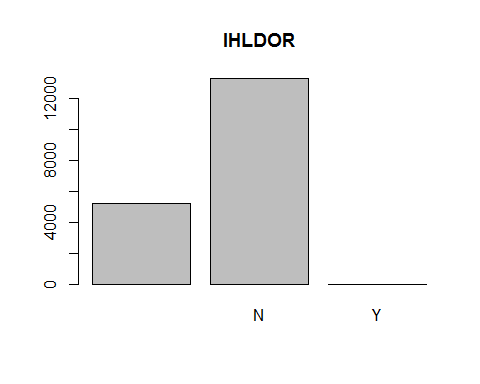
# IMINPK  
barplot(table(Inv\_final$IMINPK),main="IMINPK")



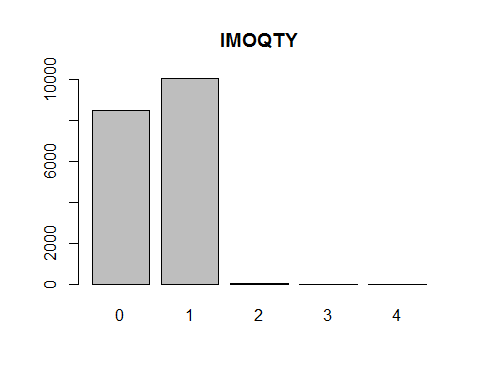
# ISTDPK  
barplot(table(Inv\_final$ISTDPK),main="ISTDPK")



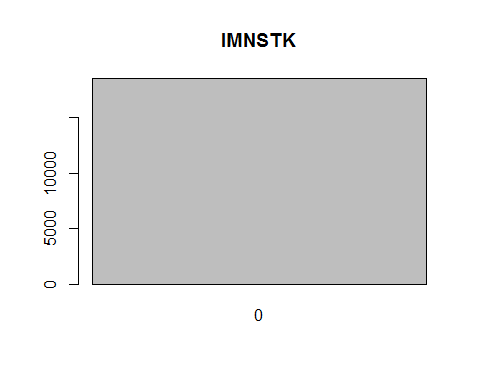
# IHLDOR - mostly empty  
barplot(table(Inv\_final$IHLDOR),main="IHLDOR")



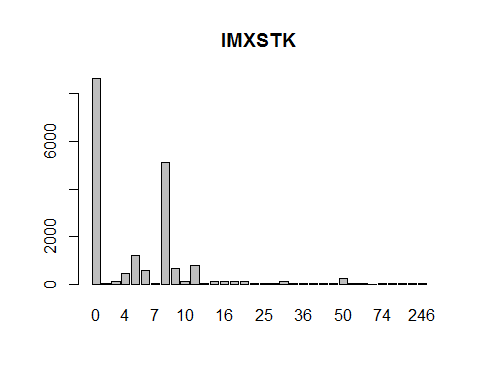
# IMOQTY - mostly empty  
barplot(table(Inv\_final$IMOQTY),main="IMOQTY")



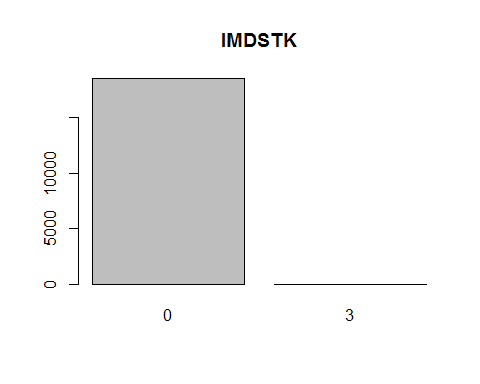
# IMNSTK - empty  
barplot(table(Inv\_final$IMNSTK),main="IMNSTK")



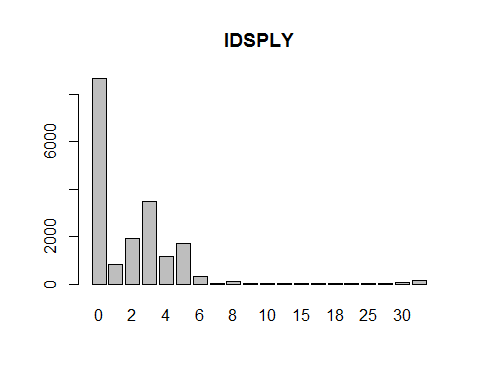
# IMXSTK  
barplot(table(Inv\_final$IMXSTK),main="IMXSTK")



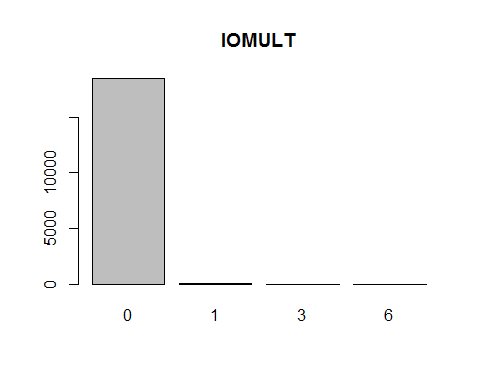
# IMDSTK - empty  
barplot(table(Inv\_final$IMDSTK),main="IMDSTK")



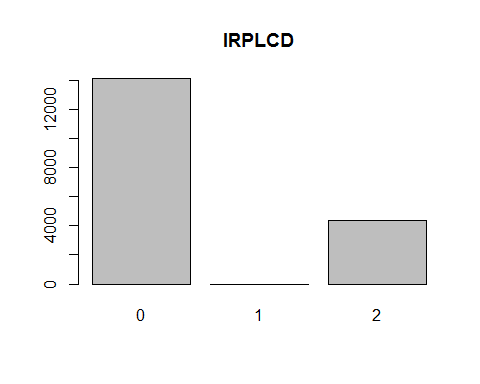
# IDSPLY  
barplot(table(Inv\_final$IDSPLY),main="IDSPLY")



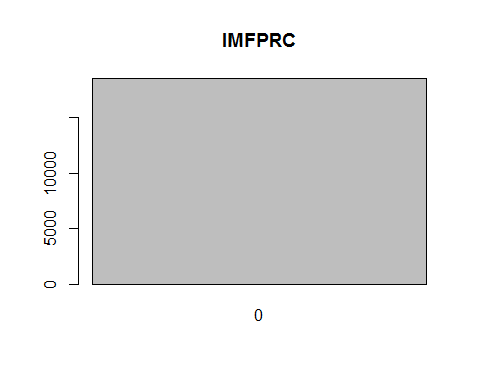
# IOMULT - mostly empty  
barplot(table(Inv\_final$IOMULT),main="IOMULT")



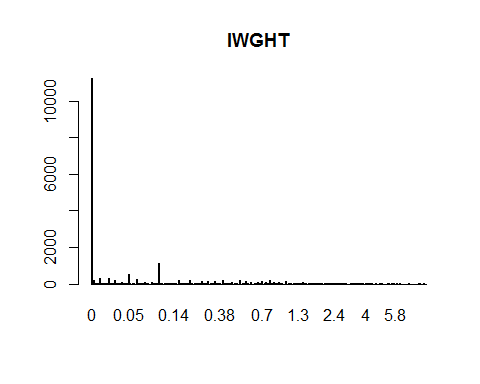
# IRPLCD - convert NA's to 0  
Inv\_final$IRPLCD[is.na(Inv\_final$IRPLCD)]<-0  
barplot(table(Inv\_final$IRPLCD),main="IRPLCD")



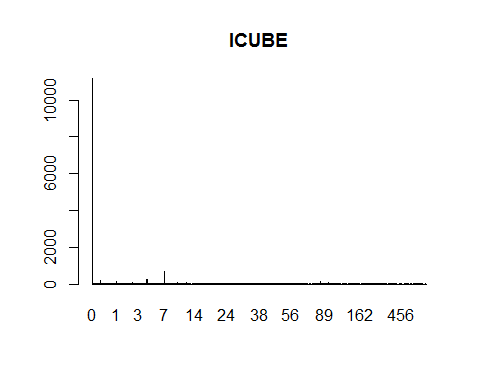
# IMFPRC - empty  
barplot(table(Inv\_final$IMFPRC),main="IMFPRC")



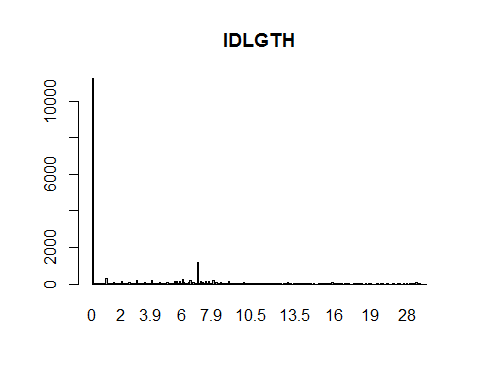
# IWGHT - there is one outlier that throws the significance off  
barplot(table(Inv\_final$IWGHT),main="IWGHT")



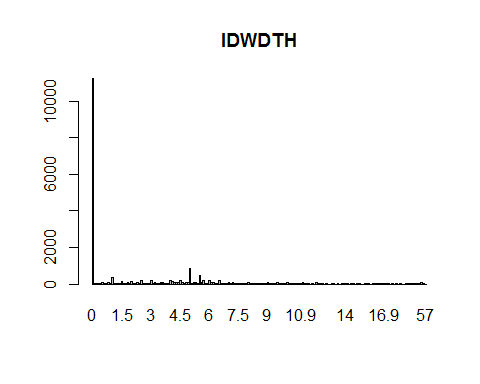
Inv\_final$IWGHT[Inv\_final$IWGHT==3001.4]=0  
# ICUBE - investigate taking the log(ICUBE)  
barplot(table(Inv\_final$ICUBE),main="ICUBE")



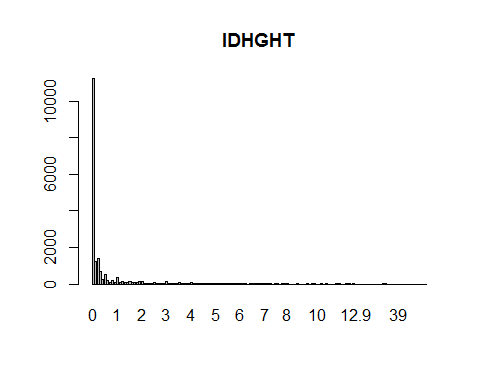
# IDLGTH - investigate taking the log(ICUBE)  
barplot(table(Inv\_final$IDLGTH),main="IDLGTH")



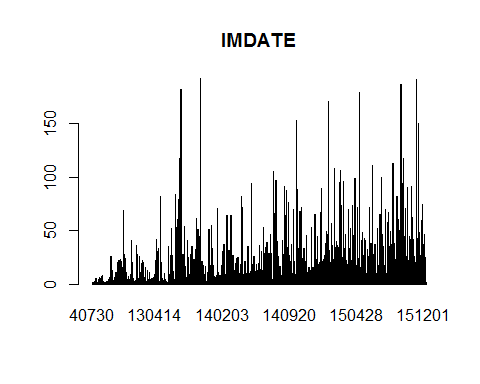
# IDWDTH - investigate taking the log(ICUBE)  
barplot(table(Inv\_final$IDWDTH),main="IDWDTH")



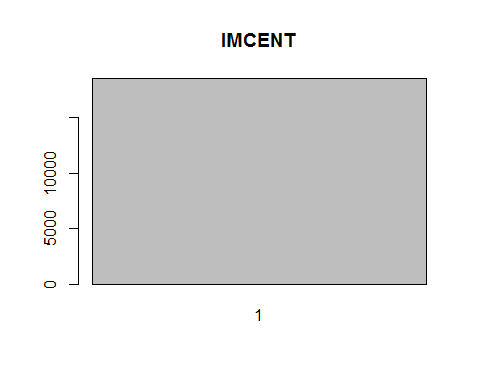
# IDHGHT - investigate taking the log(ICUBE)  
barplot(table(Inv\_final$IDHGHT),main="IDHGHT")



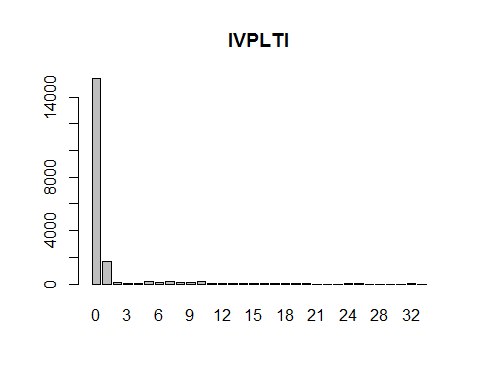
# IMDATE  
barplot(table(Inv\_final$IMDATE),main="IMDATE")



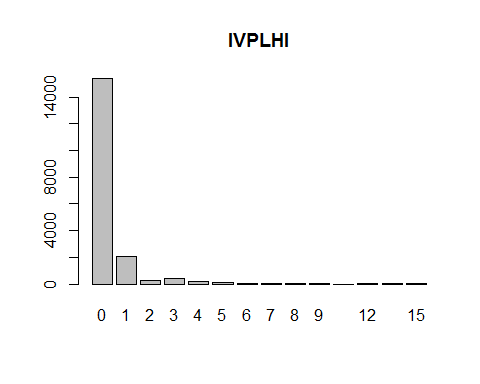
# IMCENT - empty  
barplot(table(Inv\_final$IMCENT),main="IMCENT")



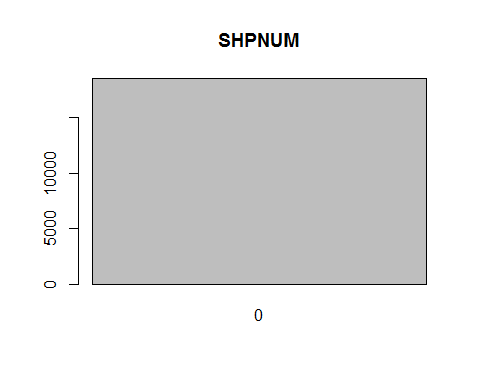
# IVPLTI - not significant  
barplot(table(Inv\_final$IVPLTI),main="IVPLTI")



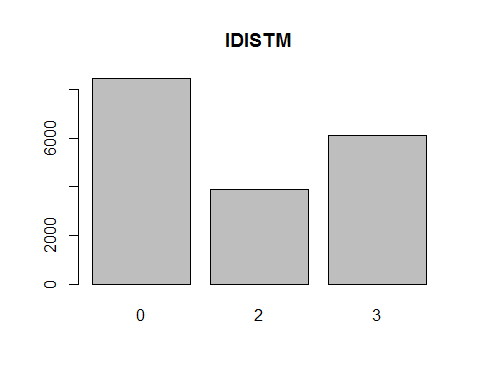
# IVPLHI  
barplot(table(Inv\_final$IVPLHI),main="IVPLHI")



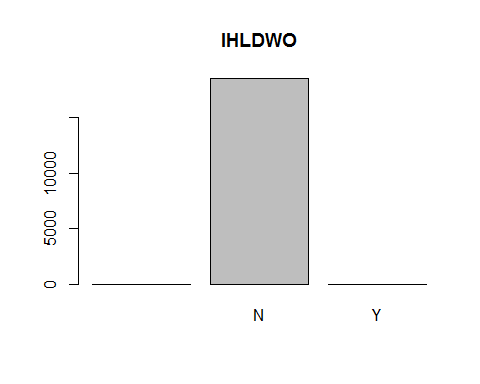
# SHPNUM - empty  
barplot(table(Inv\_final$SHPNUM),main="SHPNUM")



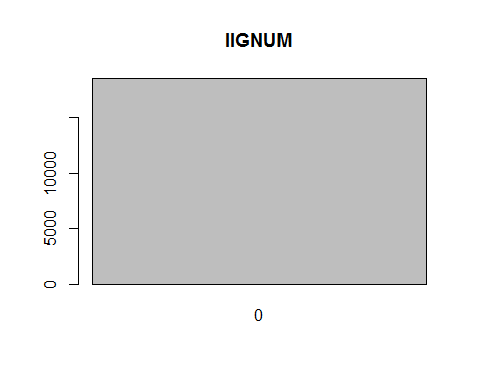
# Lots of NA's in IDISTM, replace with 0  
Inv\_final$IDISTM[is.na(Inv\_final$IDISTM)]<-0  
barplot(table(Inv\_final$IDISTM),main="IDISTM")



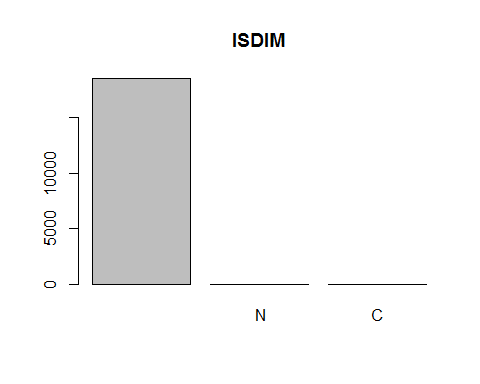
# IHLDWO - empty  
barplot(table(Inv\_final$IHLDWO),main="IHLDWO")



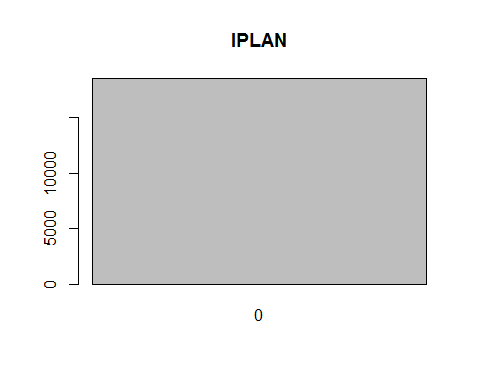
# IIGNUM - empty  
barplot(table(Inv\_final$IIGNUM),main="IIGNUM")



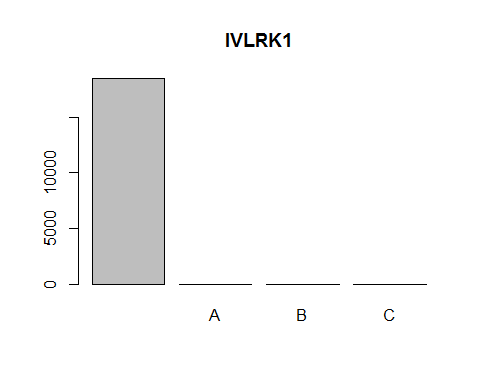
# ISDIM - empty  
barplot(table(Inv\_final$ISDIM),main="ISDIM")



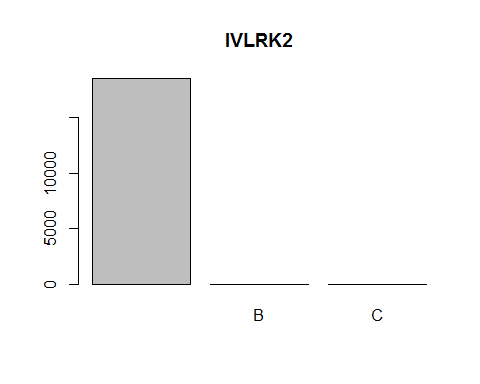
# IVATCD - all NA's  
# IPLAN - empty  
barplot(table(Inv\_final$IPLAN),main="IPLAN")



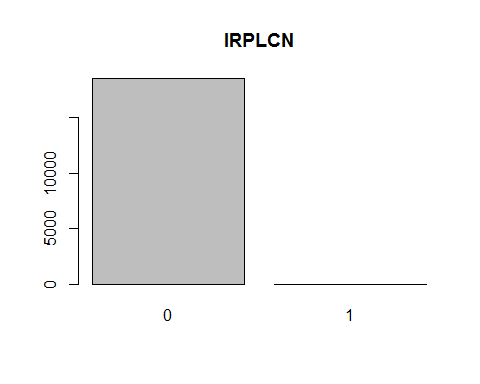
# IVLRK1 - empty  
barplot(table(Inv\_final$IVLRK1),main="IVLRK1")



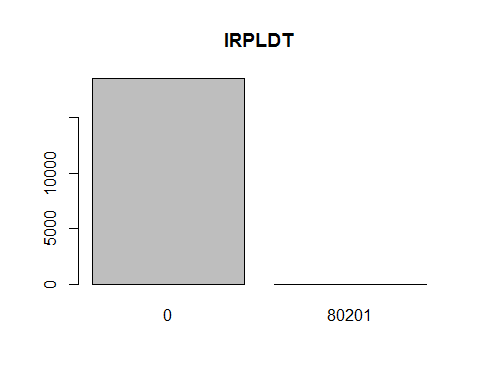
# IVLRK2 - empty  
barplot(table(Inv\_final$IVLRK2),main="IVLRK2")



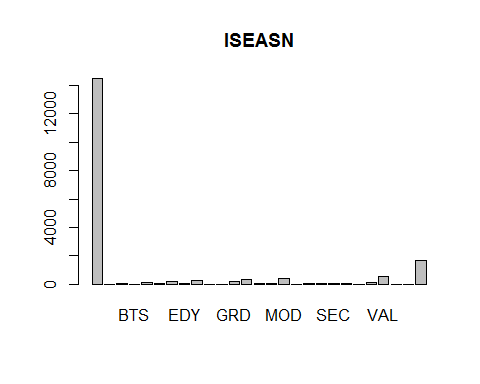
# IVLRK3 - all NA's  
# IVLRK4 - all NA's  
# IRPLCN - empty  
barplot(table(Inv\_final$IRPLCN),main="IRPLCN")



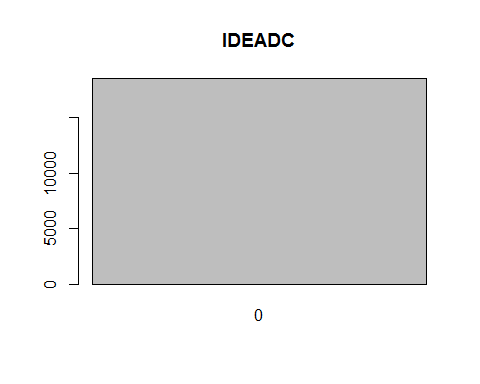
# IRPLDT - empty  
barplot(table(Inv\_final$IRPLDT),main="IRPLDT")



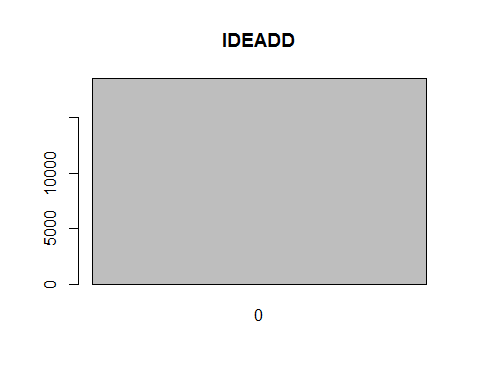
# ISEASN  
barplot(table(Inv\_final$ISEASN),main="ISEASN")



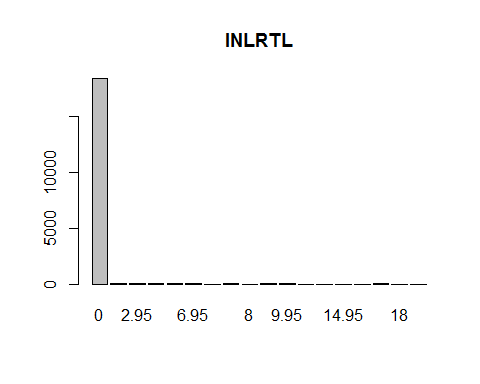
# IDEADC  
barplot(table(Inv\_final$IDEADC),main="IDEADC")



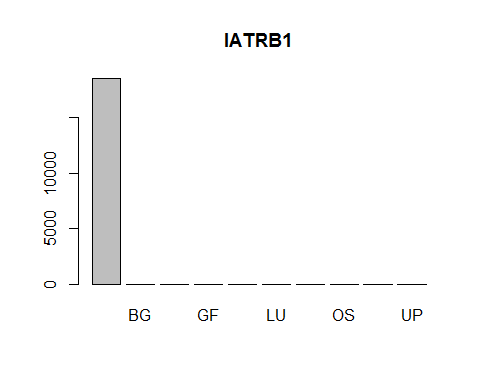
# IDEADD - empty  
barplot(table(Inv\_final$IDEADD),main="IDEADD")



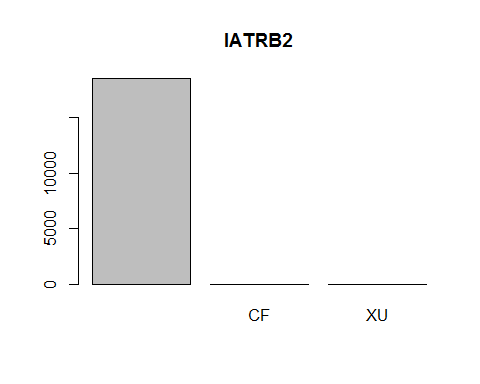
# INLRTL  
barplot(table(Inv\_final$INLRTL),main="INLRTL")



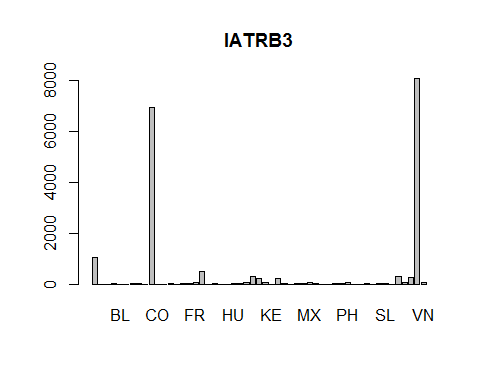
# IHANDL is all NA's  
# IATRB1 - empty  
barplot(table(Inv\_final$IATRB1),main="IATRB1")



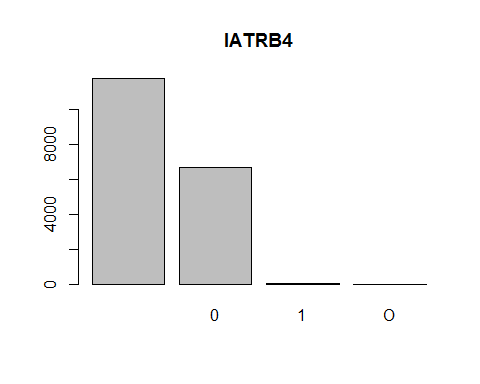
# IATRB2 - empty  
barplot(table(Inv\_final$IATRB2),main="IATRB2")



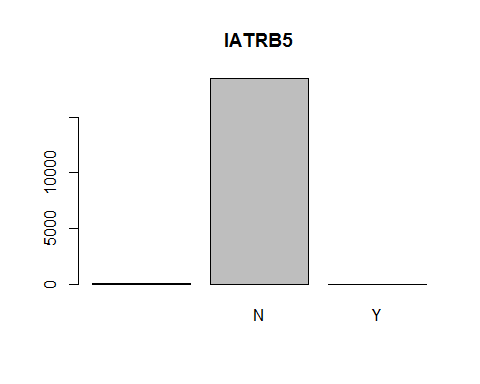
# IATRB3  
barplot(table(Inv\_final$IATRB3),main="IATRB3")



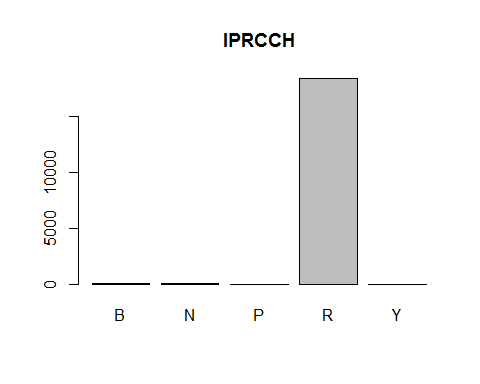
# IATRB4 - empty  
barplot(table(Inv\_final$IATRB4),main="IATRB4")



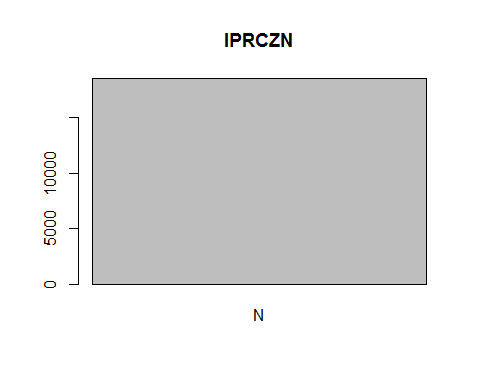
# IATRB5 - all N's  
barplot(table(Inv\_final$IATRB5),main="IATRB5")



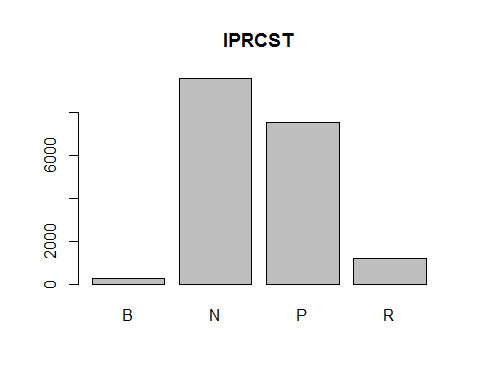
# IPRCCH  
barplot(table(Inv\_final$IPRCCH),main="IPRCCH")



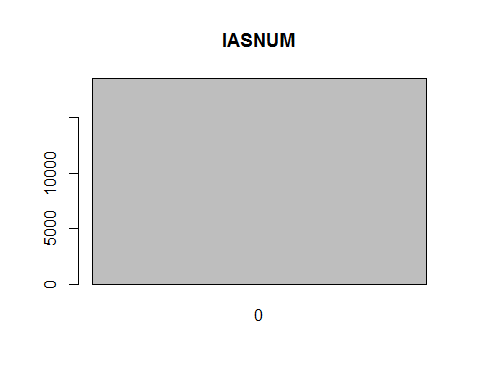
# IPRCZN - all N's  
barplot(table(Inv\_final$IPRCZN),main="IPRCZN")



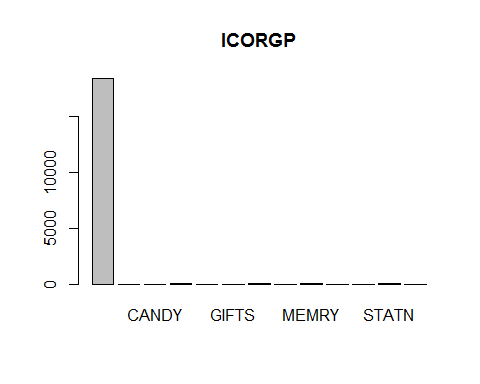
# IPRCST  
barplot(table(Inv\_final$IPRCST),main="IPRCST")



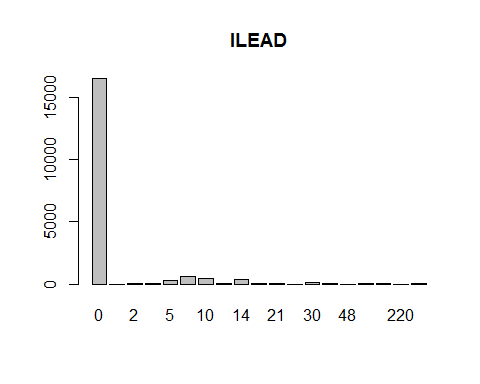
# IASNUM - empty  
barplot(table(Inv\_final$IASNUM),main="IASNUM")



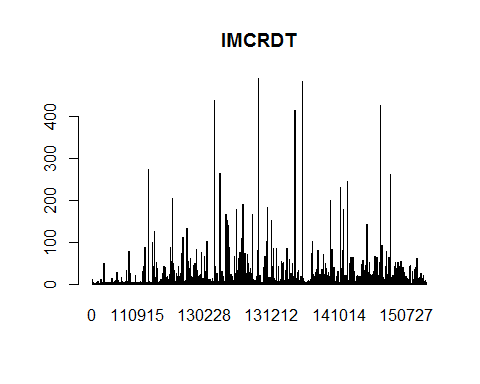
# ICORGP  
barplot(table(Inv\_final$ICORGP),main="ICORGP")



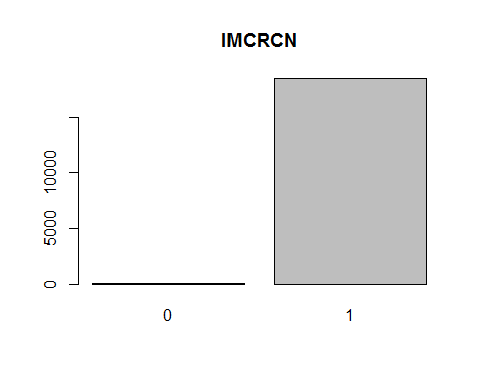
# ILEAD  
barplot(table(Inv\_final$ILEAD),main="ILEAD")



# IHZCOD - all NA's  
# IFRACT - all NA's  
# IMCRDT - distributed  
barplot(table(Inv\_final$IMCRDT),main="IMCRDT")

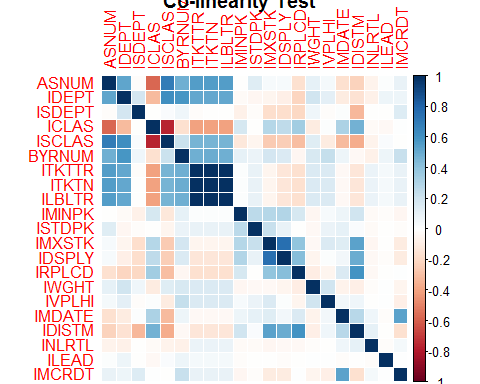


# IMCRCN - all 1's  
barplot(table(Inv\_final$IMCRCN),main="IMCRCN")



# Check cross correlation of attributes to eliminate reduncancy

Create corrplot of Inventory Master (we see that ITKTTR, ITKTN and ILBLTR are highly correlated. ITKTTR, it turns out is not a significant predictor of sales, so it will be removed. ILBLTR will also be removed due to colinearity.)



# Build basis Linear Regression model

Linear Regression model (on Sales and Inventory only). R-squared: 0.3234

##   
## Call:  
## lm(formula = IBWK01\_sum ~ IBHAND\_sum + IBWK02\_sum + IBWK03\_sum +   
## IBWK04\_sum + IBWK05\_sum + IBWK06\_sum + IBWK07\_sum + IBWK08\_sum,   
## data = Inv\_final)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.932 -0.073 -0.060 0.003 85.439   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0726392 0.0086790 8.370 < 2e-16 \*\*\*  
## IBHAND\_sum -0.0126419 0.0004206 -30.058 < 2e-16 \*\*\*  
## IBWK02\_sum 0.3861978 0.0131358 29.401 < 2e-16 \*\*\*  
## IBWK03\_sum 0.4871922 0.0170954 28.498 < 2e-16 \*\*\*  
## IBWK04\_sum 0.3668130 0.0122928 29.840 < 2e-16 \*\*\*  
## IBWK05\_sum 0.0614428 0.0166662 3.687 0.000228 \*\*\*  
## IBWK06\_sum 0.3215253 0.0173860 18.493 < 2e-16 \*\*\*  
## IBWK07\_sum 0.3214523 0.0207956 15.458 < 2e-16 \*\*\*  
## IBWK08\_sum 0.0354741 0.0157216 2.256 0.024057 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.089 on 18484 degrees of freedom  
## Multiple R-squared: 0.3234, Adjusted R-squared: 0.3231   
## F-statistic: 1104 on 8 and 18484 DF, p-value: < 2.2e-16

Complete working Linear Regression model (all non-empty attributes). R-squared: 0.4574

##   
## Call:  
## lm(formula = IBWK01\_sum ~ IBHAND\_sum + IBWK02\_sum + IBWK03\_sum +   
## IBWK04\_sum + IBWK05\_sum + IBWK06\_sum + IBWK07\_sum + IBWK08\_sum +   
## ASNUM + IDEPT + ISDEPT + ICLAS + ISCLAS + IMFGNO + BYRNUM +   
## ITKTTR + ITKTN + ILBLTR + IFINLN + ISTYPE + IMINPK + ISTDPK +   
## IMXSTK + IDSPLY + IRPLCD + IWGHT + IVPLHI + IMDATE + IDISTM +   
## ISEASN + INLRTL + IATRB3 + IPRCCH + IPRCST + ICORGP + ILEAD +   
## IMCRDT, data = Inv\_final)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30.243 -0.129 -0.012 0.080 60.455   
##   
## Coefficients: (5 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.817e-02 3.133e-01 -0.250 0.802946   
## IBHAND\_sum -1.037e-02 4.251e-04 -24.400 < 2e-16 \*\*\*  
## IBWK02\_sum 3.172e-01 1.238e-02 25.614 < 2e-16 \*\*\*  
## IBWK03\_sum 4.436e-01 1.564e-02 28.361 < 2e-16 \*\*\*  
## IBWK04\_sum 2.882e-01 1.134e-02 25.410 < 2e-16 \*\*\*  
## IBWK05\_sum 1.071e-01 1.521e-02 7.039 2.01e-12 \*\*\*  
## IBWK06\_sum 2.718e-01 1.587e-02 17.124 < 2e-16 \*\*\*  
## IBWK07\_sum 2.530e-01 1.911e-02 13.236 < 2e-16 \*\*\*  
## IBWK08\_sum 3.744e-02 1.441e-02 2.598 0.009383 \*\*   
## ASNUM -2.418e-07 1.081e-07 -2.237 0.025298 \*   
## IDEPT 4.405e-04 1.833e-04 2.403 0.016265 \*   
## ISDEPT -2.459e-03 1.191e-03 -2.064 0.038992 \*   
## ICLAS -1.076e-03 3.790e-04 -2.839 0.004525 \*\*   
## ISCLAS -1.184e-03 5.268e-04 -2.247 0.024643 \*   
## IMFGNO0147 7.305e-02 3.111e-01 0.235 0.814360   
## IMFGNO0307 8.077e-03 1.973e-01 0.041 0.967344   
## IMFGNO0397 9.522e-02 2.730e-01 0.349 0.727280   
## IMFGNO0467 2.445e-01 1.003e+00 0.244 0.807442   
## IMFGNO0473 2.749e-02 5.871e-01 0.047 0.962646   
## IMFGNO0753 -1.215e-01 7.103e-01 -0.171 0.864153   
## IMFGNO1003730 -1.278e-02 1.786e-01 -0.072 0.942969   
## IMFGNO1042 -1.097e-01 3.227e-01 -0.340 0.733811   
## IMFGNO1043 -5.646e-02 1.689e-01 -0.334 0.738137   
## IMFGNO1047 -4.153e-02 1.790e-01 -0.232 0.816517   
## IMFGNO1127 4.346e-01 9.939e-01 0.437 0.661917   
## IMFGNO1413 -4.258e-02 1.845e-01 -0.231 0.817536   
## IMFGNO1417 2.444e-01 1.062e+00 0.230 0.818070   
## IMFGNO1762 -3.292e-02 9.940e-01 -0.033 0.973576   
## IMFGNO1875 6.593e-02 9.952e-01 0.066 0.947181   
## IMFGNO2113 -4.469e-02 9.939e-01 -0.045 0.964133   
## IMFGNO2125 1.954e-01 9.957e-01 0.196 0.844385   
## IMFGNO2407 -5.704e-01 9.953e-01 -0.573 0.566539   
## IMFGNO2417 -5.057e-01 4.143e-01 -1.221 0.222194   
## IMFGNO2467 -2.370e-01 9.948e-01 -0.238 0.811712   
## IMFGNO2497 -2.952e-01 1.272e+00 -0.232 0.816485   
## IMFGNO2607 -5.499e-02 7.131e-01 -0.077 0.938535   
## IMFGNO2657 -1.328e-02 9.953e-01 -0.013 0.989354   
## IMFGNO2803 -1.563e-01 9.949e-01 -0.157 0.875154   
## IMFGNO4132 -7.716e-02 2.665e-01 -0.290 0.772147   
## IMFGNO4137 -5.486e-02 3.390e-01 -0.162 0.871446   
## IMFGNO4157 -3.825e-03 1.008e+00 -0.004 0.996974   
## IMFGNO4227 -3.944e-01 1.394e+00 -0.283 0.777335   
## IMFGNO4237 2.332e-02 5.025e-01 0.046 0.962982   
## IMFGNO4247 3.167e-02 1.011e+00 0.031 0.975010   
## IMFGNO4363 -3.138e-01 1.204e+00 -0.261 0.794461   
## IMFGNO4373 -3.581e-01 1.078e+00 -0.332 0.739848   
## IMFGNO4772 -7.302e-01 2.324e-01 -3.143 0.001677 \*\*   
## IMFGNO4773 -7.004e-01 2.122e-01 -3.300 0.000967 \*\*\*  
## IMFGNO4777 -4.389e-01 1.901e-01 -2.309 0.020970 \*   
## IMFGNO4843 -6.587e-01 4.352e-01 -1.514 0.130128   
## IMFGNO4867 -5.624e-01 9.959e-01 -0.565 0.572302   
## IMFGNO4977 -3.628e-01 2.440e-01 -1.487 0.137134   
## IMFGNO5107 -7.076e-01 4.689e-01 -1.509 0.131312   
## IMFGNO5457 -3.123e-01 1.006e+00 -0.310 0.756275   
## IMFGNO5691 -2.164e-01 5.083e-01 -0.426 0.670321   
## IMFGNO5692 -1.457e-01 4.777e-01 -0.305 0.760298   
## IMFGNO5693 -1.584e-01 2.327e-01 -0.681 0.495978   
## IMFGNO5697 -1.809e-01 2.120e-01 -0.853 0.393721   
## IMFGNO5712 -1.296e-01 9.991e-01 -0.130 0.896751   
## IMFGNO5743 -1.074e-01 3.775e-01 -0.284 0.776080   
## IMFGNO6147 1.079e-01 4.052e-01 0.266 0.790058   
## IMFGNO6337 -4.481e-02 3.352e-01 -0.134 0.893637   
## IMFGNO6413 -1.116e-01 3.584e-01 -0.311 0.755576   
## IMFGNO6417 -1.180e-01 5.308e-01 -0.222 0.824113   
## IMFGNO6557 4.019e-02 7.331e-01 0.055 0.956282   
## IMFGNO6577 1.142e-01 6.521e-01 0.175 0.860994   
## IMFGNO6647 -5.662e-02 4.826e-01 -0.117 0.906600   
## IMFGNO6893 -6.654e-02 9.976e-01 -0.067 0.946823   
## IMFGNO6897 -1.631e-01 2.375e-01 -0.687 0.492189   
## IMFGNO6903 -1.903e-01 9.978e-01 -0.191 0.848729   
## IMFGNO7172 -8.813e-02 4.238e-01 -0.208 0.835280   
## IMFGNO7173 -6.156e-02 3.256e-01 -0.189 0.850060   
## IMFGNO7177 -1.013e-01 2.787e-01 -0.363 0.716238   
## IMFGNO7247 -4.940e-02 1.991e-01 -0.248 0.804029   
## IMFGNO7493 -1.291e-01 7.222e-01 -0.179 0.858087   
## IMFGNO7657 -2.089e-01 1.003e+00 -0.208 0.835090   
## IMFGNO8811 -1.565e-01 9.951e-01 -0.157 0.875067   
## IMFGNO9011 -8.774e-02 1.800e-01 -0.487 0.626012   
## IMFGNO9029 -1.385e-01 1.546e-01 -0.895 0.370560   
## IMFGNO9041 -8.504e-02 3.228e-01 -0.263 0.792197   
## IMFGNO9059 -5.111e-02 9.935e-01 -0.051 0.958971   
## IMFGNO9069 -1.989e-01 2.100e-01 -0.947 0.343600   
## IMFGNO9079 5.698e-03 1.729e-01 0.033 0.973716   
## IMFGNO9089 -7.323e-02 1.731e-01 -0.423 0.672175   
## IMFGNO9099 1.764e-01 2.090e-01 0.844 0.398573   
## IMFGNO9109 -2.400e-01 1.810e-01 -1.326 0.184871   
## IMFGNO9119 -2.458e-01 2.063e-01 -1.191 0.233583   
## IMFGNO9139 -2.029e-01 1.877e-01 -1.081 0.279610   
## IMFGNO9269 -5.524e-02 1.655e-01 -0.334 0.738519   
## IMFGNO9291 -1.614e-01 7.489e-01 -0.215 0.829392   
## IMFGNO9311 -1.936e-01 5.171e-01 -0.374 0.708109   
## IMFGNO9331 -7.655e-02 2.897e-01 -0.264 0.791606   
## IMFGNO9341 -6.989e-01 1.009e+00 -0.693 0.488350   
## IMFGNO9361 -4.059e-01 2.204e-01 -1.842 0.065522 .   
## IMFGNO9381 -1.207e-01 2.248e-01 -0.537 0.591294   
## IMFGNO9391 -1.054e-01 2.820e-01 -0.374 0.708499   
## IMFGNO9411 -7.614e-02 2.867e-01 -0.266 0.790564   
## IMFGNO9421 -7.857e-02 2.930e-01 -0.268 0.788568   
## IMFGNO9431 -9.871e-02 3.836e-01 -0.257 0.796949   
## IMFGNO9451 -3.121e-01 2.555e-01 -1.221 0.221915   
## IMFGNO9471 -1.702e-01 5.900e-01 -0.289 0.772944   
## IMFGNO9499 -1.505e-01 2.446e-01 -0.615 0.538376   
## IMFGNO9509 -7.972e-01 9.903e-01 -0.805 0.420829   
## IMFGNO9529 7.888e-02 1.797e-01 0.439 0.660749   
## IMFGNO9539 6.486e-03 8.016e-01 0.008 0.993544   
## IMFGNO9549 -9.054e-02 2.031e-01 -0.446 0.655814   
## IMFGNO9559 -1.052e-01 2.188e-01 -0.481 0.630690   
## IMFGNO9569 -1.738e-01 3.265e-01 -0.532 0.594489   
## IMFGNO9579 -6.145e-02 1.880e-01 -0.327 0.743796   
## IMFGNO9589 -8.624e-02 2.181e-01 -0.395 0.692486   
## IMFGNO9599 -7.869e-02 2.173e-01 -0.362 0.717299   
## IMFGNO9629 -1.018e-01 4.308e-01 -0.236 0.813186   
## IMFGNO9649 3.404e-02 2.992e-01 0.114 0.909419   
## IMFGNO9659 5.615e-02 6.721e-01 0.084 0.933418   
## IMFGNO9669 3.583e-02 2.806e-01 0.128 0.898412   
## IMFGNO9679 2.782e-01 1.802e-01 1.544 0.122528   
## IMFGNO9689 4.004e-02 3.023e-01 0.132 0.894633   
## IMFGNO9699 -1.243e-01 3.433e-01 -0.362 0.717423   
## IMFGNO9739 3.653e-01 1.845e-01 1.980 0.047724 \*   
## IMFGNO9749 -1.202e-01 3.642e-01 -0.330 0.741457   
## IMFGNO9769 1.605e-01 5.471e-01 0.293 0.769257   
## IMFGNO9789 3.891e-02 5.349e-01 0.073 0.942015   
## IMFGNOANNIVERSARY 1.155e-01 6.945e-01 0.166 0.867964   
## IMFGNOBATH -4.949e-01 2.867e-01 -1.726 0.084299 .   
## IMFGNOE2049 CRYS/M -7.054e-02 4.394e-01 -0.161 0.872456   
## IMFGNOMEMO FOLIO 9.773e-03 9.818e-01 0.010 0.992058   
## IMFGNOMSL20 -2.427e-01 9.824e-01 -0.247 0.804845   
## IMFGNOPORRINGER 1.217e-01 9.828e-01 0.124 0.901426   
## IMFGNOPZ-RABBIT -5.189e-03 4.399e-01 -0.012 0.990589   
## IMFGNOWALLET 1.016e-01 9.816e-01 0.104 0.917544   
## IMFGNOWHITE CITY 5.649e-03 4.917e-01 0.011 0.990834   
## IMFGNOY -4.474e-02 1.727e-01 -0.259 0.795551   
## BYRNUM -1.659e-02 6.295e-03 -2.636 0.008401 \*\*   
## ITKTTR 9.132e-02 3.559e-01 0.257 0.797469   
## ITKTN 1.222e+00 4.044e-01 3.021 0.002522 \*\*   
## ILBLTR -1.369e+00 5.386e-01 -2.542 0.011015 \*   
## IFINLNACCES 3.238e-01 1.077e+00 0.301 0.763593   
## IFINLNCARDS -8.106e-02 1.051e-01 -0.771 0.440422   
## IFINLNCOLLE 8.333e-01 1.751e+00 0.476 0.634252   
## IFINLNDISC 6.765e-01 1.066e+00 0.635 0.525727   
## IFINLNGIFT -2.152e-01 1.135e+00 -0.190 0.849642   
## IFINLNOTHER -1.424e+00 3.873e-01 -3.676 0.000237 \*\*\*  
## IFINLNSTATN -2.932e-03 4.619e-01 -0.006 0.994935   
## IFINLNWRAP 1.081e-02 6.969e-01 0.016 0.987625   
## ISTYPE05 -1.321e+00 5.126e-01 -2.578 0.009952 \*\*   
## ISTYPECM -4.313e-02 2.704e-01 -0.160 0.873269   
## ISTYPEDM -7.438e-02 2.225e-01 -0.334 0.738201   
## ISTYPEGW 3.265e+01 7.338e-01 44.496 < 2e-16 \*\*\*  
## ISTYPEIM -2.279e-02 2.245e-01 -0.102 0.919151   
## IMINPK 7.474e-03 1.589e-03 4.704 2.57e-06 \*\*\*  
## ISTDPK -1.891e-04 7.062e-05 -2.677 0.007430 \*\*   
## IMXSTK 1.843e-03 1.452e-03 1.269 0.204401   
## IDSPLY 6.272e-03 2.920e-03 2.148 0.031703 \*   
## IRPLCD -7.418e-03 1.199e-02 -0.619 0.536121   
## IWGHT -1.263e-02 1.097e-02 -1.152 0.249540   
## IVPLHI -2.370e-02 8.376e-03 -2.830 0.004660 \*\*   
## IMDATE 4.973e-06 1.263e-06 3.938 8.23e-05 \*\*\*  
## IDISTM -5.936e-02 1.187e-02 -5.000 5.78e-07 \*\*\*  
## ISEASNBOS NA NA NA NA   
## ISEASNCAL 1.809e-01 1.528e-01 1.184 0.236289   
## ISEASNCOM NA NA NA NA   
## ISEASNEAS -7.800e-02 1.363e-01 -0.572 0.567242   
## ISEASNEDY -8.119e-02 1.050e+00 -0.077 0.938387   
## ISEASNFAD -8.631e-02 1.475e-01 -0.585 0.558428   
## ISEASNGRD -6.899e-04 1.496e-01 -0.005 0.996320   
## ISEASNHAL -1.054e-01 1.906e-01 -0.553 0.580052   
## ISEASNHNK 5.490e-01 1.730e-01 3.174 0.001507 \*\*   
## ISEASNJNY -1.751e-02 2.366e-01 -0.074 0.940992   
## ISEASNMOD -7.868e-02 1.116e-01 -0.705 0.480757   
## ISEASNNYD 8.547e-02 6.947e-01 0.123 0.902076   
## ISEASNPAS -1.958e-01 5.689e-01 -0.344 0.730737   
## ISEASNSEC NA NA NA NA   
## ISEASNSTP NA NA NA NA   
## ISEASNTHG 2.369e-01 9.951e-01 0.238 0.811798   
## ISEASNVAL -1.566e-02 1.172e-01 -0.134 0.893682   
## ISEASNXMS 6.331e-01 7.860e-02 8.055 8.43e-16 \*\*\*  
## INLRTL 2.331e-02 1.309e-02 1.780 0.075063 .   
## IATRB3AU 4.610e-01 3.290e-01 1.401 0.161217   
## IATRB3CA 1.585e-02 2.002e-01 0.079 0.936906   
## IATRB3CH -1.918e-01 1.627e-01 -1.178 0.238685   
## IATRB3CN 3.297e-02 3.731e-02 0.884 0.376844   
## IATRB3DE 4.188e-01 3.305e-01 1.267 0.205110   
## IATRB3EN 3.085e-02 6.954e-01 0.044 0.964619   
## IATRB3ES 2.024e-02 1.702e-01 0.119 0.905336   
## IATRB3FR 4.639e-02 1.209e-01 0.384 0.701189   
## IATRB3GB -9.368e-02 6.435e-02 -1.456 0.145503   
## IATRB3HK -9.601e-02 3.162e-01 -0.304 0.761398   
## IATRB3HU 5.850e-02 2.288e-01 0.256 0.798209   
## IATRB3ID 5.333e-01 2.360e-01 2.260 0.023816 \*   
## IATRB3IL -1.318e-01 1.409e-01 -0.936 0.349464   
## IATRB3IN 6.497e-02 6.651e-02 0.977 0.328637   
## IATRB3IT 9.831e-02 7.481e-02 1.314 0.188815   
## IATRB3JP -3.173e-01 1.665e-01 -1.905 0.056771 .   
## IATRB3KR 1.178e-01 7.614e-02 1.548 0.121711   
## IATRB3LB -3.297e-02 1.954e-01 -0.169 0.866043   
## IATRB3MA 1.089e-02 5.677e-01 0.019 0.984693   
## IATRB3MG 8.347e-02 2.345e-01 0.356 0.721916   
## IATRB3MX -1.844e-03 1.395e-01 -0.013 0.989452   
## IATRB3MY 1.805e-01 3.038e-01 0.594 0.552411   
## IATRB3NP -2.332e-02 2.855e-01 -0.082 0.934919   
## IATRB3PE -1.319e-01 2.745e-01 -0.480 0.630883   
## IATRB3PH 1.627e-01 1.361e-01 1.196 0.231703   
## IATRB3PT -4.002e-01 5.693e-01 -0.703 0.482085   
## IATRB3SG 1.678e-01 5.688e-01 0.295 0.767994   
## IATRB3SL 2.935e+00 6.947e-01 4.224 2.41e-05 \*\*\*  
## IATRB3TH -1.094e-01 6.793e-02 -1.610 0.107385   
## IATRB3TR -5.526e-02 1.456e-01 -0.380 0.704249   
## IATRB3TW 5.151e-01 8.147e-02 6.322 2.64e-10 \*\*\*  
## IATRB3US 2.675e-02 3.700e-02 0.723 0.469780   
## IATRB3VN -1.159e-01 1.642e-01 -0.706 0.480110   
## IPRCCHN -1.737e+00 2.118e-01 -8.199 2.59e-16 \*\*\*  
## IPRCCHR -9.299e-02 1.333e-01 -0.698 0.485412   
## IPRCSTN 1.572e-01 6.947e-02 2.263 0.023632 \*   
## IPRCSTP 8.668e-02 6.890e-02 1.258 0.208360   
## IPRCSTR 1.323e-01 7.434e-02 1.780 0.075054 .   
## ICORGPCARDS 4.222e-02 3.787e-01 0.111 0.911227   
## ICORGPCOLLE -7.822e-01 1.449e+00 -0.540 0.589259   
## ICORGPGIFTS 3.971e-02 1.260e+00 0.032 0.974863   
## ICORGPHOMDE -2.666e-01 1.138e+00 -0.234 0.814794   
## ICORGPMEMRY 1.111e-01 9.819e-01 0.113 0.909918   
## ICORGPSEASN 1.978e+01 1.525e+00 12.965 < 2e-16 \*\*\*  
## ICORGPSTATN 1.606e-01 4.158e-01 0.386 0.699371   
## ICORGPWRAP NA NA NA NA   
## ILEAD -3.921e-04 6.231e-04 -0.629 0.529130   
## IMCRDT -2.300e-06 1.030e-06 -2.234 0.025523 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.981 on 18273 degrees of freedom  
## Multiple R-squared: 0.4574, Adjusted R-squared: 0.4509   
## F-statistic: 70.33 on 219 and 18273 DF, p-value: < 2.2e-16

Add top 50 words to Linear Regression model. R-squared: 0.4586

##   
## Call:  
## lm(formula = IBWK01\_sum ~ IBHAND\_sum + IBWK02\_sum + IBWK03\_sum +   
## IBWK04\_sum + IBWK05\_sum + IBWK06\_sum + IBWK07\_sum + IBWK08\_sum +   
## ASNUM + IDEPT + ISDEPT + ICLAS + ISCLAS + IMFGNO + BYRNUM +   
## ITKTTR + ITKTN + ILBLTR + IFINLN + ISTYPE + IMINPK + ISTDPK +   
## IMXSTK + IDSPLY + IRPLCD + IWGHT + IVPLHI + IMDATE + IDISTM +   
## ISEASN + INLRTL + IATRB3 + IPRCCH + IPRCST + ICORGP + ILEAD +   
## IMCRDT + w\_babi + w\_bag + w\_bdi + w\_bird + w\_birthday + w\_black +   
## w\_blue + w\_box + w\_bracelet + w\_butterfli + w\_cake + w\_candl +   
## w\_card + w\_cardx + w\_conv + w\_crystal + w\_dog + w\_dress +   
## w\_earring + w\_floral + w\_flower + w\_gem + w\_general + w\_girl +   
## w\_glass + w\_glitter + w\_gold + w\_handmad + w\_happi + w\_heart +   
## w\_love + w\_med + w\_mini + w\_mom + w\_neck + w\_necklac + w\_note +   
## w\_pink + w\_print + w\_red + w\_ribbon + w\_set + w\_silver +   
## w\_soap + w\_thank + w\_tree + w\_vintag + w\_wed + w\_white +   
## w\_xbc, data = Inv\_final)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30.260 -0.146 -0.011 0.086 60.455   
##   
## Coefficients: (5 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.456e-01 3.163e-01 -0.460 0.645236   
## IBHAND\_sum -1.041e-02 4.263e-04 -24.418 < 2e-16 \*\*\*  
## IBWK02\_sum 3.171e-01 1.240e-02 25.567 < 2e-16 \*\*\*  
## IBWK03\_sum 4.425e-01 1.566e-02 28.248 < 2e-16 \*\*\*  
## IBWK04\_sum 2.886e-01 1.136e-02 25.406 < 2e-16 \*\*\*  
## IBWK05\_sum 1.059e-01 1.524e-02 6.951 3.74e-12 \*\*\*  
## IBWK06\_sum 2.718e-01 1.590e-02 17.098 < 2e-16 \*\*\*  
## IBWK07\_sum 2.524e-01 1.916e-02 13.172 < 2e-16 \*\*\*  
## IBWK08\_sum 3.794e-02 1.444e-02 2.628 0.008601 \*\*   
## ASNUM -2.544e-07 1.094e-07 -2.326 0.020035 \*   
## IDEPT 4.103e-04 1.888e-04 2.173 0.029799 \*   
## ISDEPT -2.234e-03 1.323e-03 -1.689 0.091241 .   
## ICLAS -1.039e-03 3.826e-04 -2.715 0.006637 \*\*   
## ISCLAS -1.199e-03 5.361e-04 -2.236 0.025369 \*   
## IMFGNO0147 7.466e-02 3.133e-01 0.238 0.811634   
## IMFGNO0307 5.927e-02 2.147e-01 0.276 0.782521   
## IMFGNO0397 1.322e-01 2.789e-01 0.474 0.635465   
## IMFGNO0467 2.787e-01 1.007e+00 0.277 0.781879   
## IMFGNO0473 4.173e-02 5.880e-01 0.071 0.943414   
## IMFGNO0753 -9.992e-02 7.117e-01 -0.140 0.888352   
## IMFGNO1003730 3.271e-02 1.896e-01 0.173 0.863029   
## IMFGNO1042 -6.975e-02 3.295e-01 -0.212 0.832333   
## IMFGNO1043 -2.624e-02 1.758e-01 -0.149 0.881378   
## IMFGNO1047 -2.742e-02 1.811e-01 -0.151 0.879622   
## IMFGNO1127 4.845e-01 9.978e-01 0.486 0.627265   
## IMFGNO1413 -1.227e-02 1.961e-01 -0.063 0.950105   
## IMFGNO1417 2.357e-01 1.064e+00 0.222 0.824647   
## IMFGNO1762 1.280e-02 9.977e-01 0.013 0.989765   
## IMFGNO1875 5.124e-02 9.958e-01 0.051 0.958961   
## IMFGNO2113 -5.747e-02 9.944e-01 -0.058 0.953918   
## IMFGNO2125 2.707e-01 1.005e+00 0.269 0.787644   
## IMFGNO2407 -6.157e-01 9.976e-01 -0.617 0.537099   
## IMFGNO2417 -5.986e-01 4.199e-01 -1.426 0.154008   
## IMFGNO2467 -2.456e-01 9.954e-01 -0.247 0.805107   
## IMFGNO2497 -2.903e-01 1.274e+00 -0.228 0.819706   
## IMFGNO2607 -5.014e-02 7.143e-01 -0.070 0.944042   
## IMFGNO2657 -5.420e-02 9.977e-01 -0.054 0.956677   
## IMFGNO2803 -9.835e-02 9.978e-01 -0.099 0.921483   
## IMFGNO4132 -1.155e-01 2.682e-01 -0.431 0.666727   
## IMFGNO4137 -7.351e-02 3.398e-01 -0.216 0.828740   
## IMFGNO4157 -1.882e-02 1.009e+00 -0.019 0.985117   
## IMFGNO4227 -4.107e-01 1.395e+00 -0.294 0.768488   
## IMFGNO4237 3.840e-03 5.042e-01 0.008 0.993924   
## IMFGNO4247 -5.630e-02 1.013e+00 -0.056 0.955690   
## IMFGNO4363 -3.261e-01 1.205e+00 -0.271 0.786679   
## IMFGNO4373 -3.763e-01 1.079e+00 -0.349 0.727247   
## IMFGNO4772 -7.543e-01 2.346e-01 -3.215 0.001305 \*\*   
## IMFGNO4773 -7.046e-01 2.135e-01 -3.301 0.000965 \*\*\*  
## IMFGNO4777 -4.612e-01 1.915e-01 -2.408 0.016055 \*   
## IMFGNO4843 -6.827e-01 4.374e-01 -1.561 0.118533   
## IMFGNO4867 -6.210e-01 9.985e-01 -0.622 0.534042   
## IMFGNO4977 -4.007e-01 2.489e-01 -1.610 0.107455   
## IMFGNO5107 -7.239e-01 4.724e-01 -1.532 0.125480   
## IMFGNO5457 -2.887e-01 1.007e+00 -0.287 0.774396   
## IMFGNO5691 -2.072e-01 5.085e-01 -0.408 0.683634   
## IMFGNO5692 -1.433e-01 4.791e-01 -0.299 0.764875   
## IMFGNO5693 -1.573e-01 2.350e-01 -0.669 0.503380   
## IMFGNO5697 -1.917e-01 2.144e-01 -0.894 0.371109   
## IMFGNO5712 -1.160e-01 9.998e-01 -0.116 0.907655   
## IMFGNO5743 -1.067e-01 3.790e-01 -0.282 0.778263   
## IMFGNO6147 8.970e-02 4.090e-01 0.219 0.826399   
## IMFGNO6337 -6.184e-02 3.361e-01 -0.184 0.854017   
## IMFGNO6413 -1.435e-01 3.598e-01 -0.399 0.689958   
## IMFGNO6417 -2.184e-01 5.369e-01 -0.407 0.684168   
## IMFGNO6557 -6.945e-02 7.395e-01 -0.094 0.925177   
## IMFGNO6577 1.050e-02 6.591e-01 0.016 0.987286   
## IMFGNO6647 -6.349e-02 4.850e-01 -0.131 0.895850   
## IMFGNO6893 -5.936e-02 9.984e-01 -0.059 0.952594   
## IMFGNO6897 -1.765e-01 2.390e-01 -0.738 0.460284   
## IMFGNO6903 -2.290e-01 1.000e+00 -0.229 0.818945   
## IMFGNO7172 -1.281e-01 4.253e-01 -0.301 0.763296   
## IMFGNO7173 -6.942e-02 3.264e-01 -0.213 0.831584   
## IMFGNO7177 -1.092e-01 2.797e-01 -0.390 0.696307   
## IMFGNO7247 -7.856e-02 2.007e-01 -0.391 0.695522   
## IMFGNO7493 -1.383e-01 7.227e-01 -0.191 0.848210   
## IMFGNO7657 -2.263e-01 1.004e+00 -0.225 0.821668   
## IMFGNO8811 -9.055e-02 1.006e+00 -0.090 0.928252   
## IMFGNO9011 -9.739e-02 1.811e-01 -0.538 0.590768   
## IMFGNO9029 -1.534e-01 1.561e-01 -0.983 0.325789   
## IMFGNO9041 -9.467e-02 3.238e-01 -0.292 0.769995   
## IMFGNO9059 -3.243e-02 9.973e-01 -0.033 0.974058   
## IMFGNO9069 -2.034e-01 2.116e-01 -0.961 0.336478   
## IMFGNO9079 -2.390e-02 1.743e-01 -0.137 0.890927   
## IMFGNO9089 -8.633e-02 1.765e-01 -0.489 0.624813   
## IMFGNO9099 1.600e-01 2.099e-01 0.762 0.445916   
## IMFGNO9109 -2.826e-01 1.828e-01 -1.546 0.122186   
## IMFGNO9119 -2.422e-01 2.073e-01 -1.168 0.242817   
## IMFGNO9139 -2.210e-01 1.890e-01 -1.169 0.242222   
## IMFGNO9269 -7.843e-02 1.668e-01 -0.470 0.638167   
## IMFGNO9291 -1.654e-01 7.496e-01 -0.221 0.825377   
## IMFGNO9311 -2.129e-01 5.178e-01 -0.411 0.681003   
## IMFGNO9331 -9.610e-02 2.908e-01 -0.331 0.741019   
## IMFGNO9341 -7.133e-01 1.009e+00 -0.707 0.479741   
## IMFGNO9361 -4.221e-01 2.217e-01 -1.904 0.056900 .   
## IMFGNO9381 -1.250e-01 2.262e-01 -0.553 0.580524   
## IMFGNO9391 -1.312e-01 2.832e-01 -0.463 0.643024   
## IMFGNO9411 -1.194e-01 2.885e-01 -0.414 0.679093   
## IMFGNO9421 -8.439e-02 2.940e-01 -0.287 0.774096   
## IMFGNO9431 -1.085e-01 3.843e-01 -0.282 0.777683   
## IMFGNO9451 -3.300e-01 2.567e-01 -1.285 0.198672   
## IMFGNO9471 -1.809e-01 5.906e-01 -0.306 0.759317   
## IMFGNO9499 -1.780e-01 2.457e-01 -0.724 0.468937   
## IMFGNO9509 -8.509e-01 9.912e-01 -0.858 0.390675   
## IMFGNO9529 4.302e-02 1.813e-01 0.237 0.812472   
## IMFGNO9539 -2.652e-02 8.023e-01 -0.033 0.973627   
## IMFGNO9549 -9.799e-02 2.048e-01 -0.478 0.632334   
## IMFGNO9559 -1.410e-01 2.204e-01 -0.639 0.522543   
## IMFGNO9569 -2.060e-01 3.278e-01 -0.628 0.529817   
## IMFGNO9579 -7.216e-02 1.896e-01 -0.381 0.703486   
## IMFGNO9589 -9.450e-02 2.192e-01 -0.431 0.666393   
## IMFGNO9599 -8.626e-02 2.183e-01 -0.395 0.692688   
## IMFGNO9629 -1.628e-01 4.335e-01 -0.375 0.707307   
## IMFGNO9649 -4.319e-02 3.046e-01 -0.142 0.887263   
## IMFGNO9659 1.468e-02 6.733e-01 0.022 0.982606   
## IMFGNO9669 -1.335e-02 2.847e-01 -0.047 0.962601   
## IMFGNO9679 2.316e-01 1.873e-01 1.237 0.216161   
## IMFGNO9689 -3.519e-02 3.082e-01 -0.114 0.909084   
## IMFGNO9699 -1.535e-01 3.467e-01 -0.443 0.658062   
## IMFGNO9739 3.679e-01 1.877e-01 1.960 0.049980 \*   
## IMFGNO9749 -1.359e-01 3.650e-01 -0.372 0.709733   
## IMFGNO9769 1.578e-01 5.479e-01 0.288 0.773364   
## IMFGNO9789 2.990e-02 5.377e-01 0.056 0.955651   
## IMFGNOANNIVERSARY 1.017e-01 6.979e-01 0.146 0.884106   
## IMFGNOBATH -4.799e-01 2.869e-01 -1.673 0.094399 .   
## IMFGNOE2049 CRYS/M -1.155e-01 4.410e-01 -0.262 0.793464   
## IMFGNOMEMO FOLIO 1.313e-02 9.822e-01 0.013 0.989335   
## IMFGNOMSL20 -2.385e-01 9.827e-01 -0.243 0.808269   
## IMFGNOPORRINGER 1.252e-01 9.831e-01 0.127 0.898695   
## IMFGNOPZ-RABBIT 5.048e-03 4.403e-01 0.011 0.990854   
## IMFGNOWALLET 1.313e-01 9.837e-01 0.133 0.893816   
## IMFGNOWHITE CITY 4.428e-03 4.927e-01 0.009 0.992829   
## IMFGNOY -4.171e-02 1.734e-01 -0.240 0.809952   
## BYRNUM -1.785e-02 6.347e-03 -2.812 0.004926 \*\*   
## ITKTTR 1.273e-01 3.587e-01 0.355 0.722599   
## ITKTN 1.242e+00 4.049e-01 3.068 0.002156 \*\*   
## ILBLTR -1.428e+00 5.408e-01 -2.640 0.008305 \*\*   
## IFINLNACCES 3.306e-01 1.077e+00 0.307 0.758863   
## IFINLNCARDS -7.755e-02 1.061e-01 -0.731 0.464772   
## IFINLNCOLLE 8.487e-01 1.758e+00 0.483 0.629175   
## IFINLNDISC 6.199e-01 1.070e+00 0.580 0.562236   
## IFINLNGIFT -2.130e-01 1.136e+00 -0.188 0.851240   
## IFINLNOTHER -1.469e+00 3.907e-01 -3.760 0.000171 \*\*\*  
## IFINLNSTATN -5.552e-02 4.641e-01 -0.120 0.904775   
## IFINLNWRAP -1.871e-02 6.989e-01 -0.027 0.978642   
## ISTYPE05 -1.229e+00 5.213e-01 -2.358 0.018391 \*   
## ISTYPECM -1.057e-02 2.724e-01 -0.039 0.969039   
## ISTYPEDM -4.371e-02 2.241e-01 -0.195 0.845358   
## ISTYPEGW 3.274e+01 7.361e-01 44.476 < 2e-16 \*\*\*  
## ISTYPEIM -9.131e-03 2.258e-01 -0.040 0.967752   
## IMINPK 7.493e-03 1.601e-03 4.679 2.90e-06 \*\*\*  
## ISTDPK -2.184e-04 7.235e-05 -3.018 0.002545 \*\*   
## IMXSTK 1.879e-03 1.461e-03 1.286 0.198486   
## IDSPLY 8.026e-03 3.839e-03 2.090 0.036592 \*   
## IRPLCD -7.753e-03 1.211e-02 -0.640 0.522047   
## IWGHT -1.221e-02 1.102e-02 -1.107 0.268102   
## IVPLHI -2.573e-02 8.549e-03 -3.010 0.002614 \*\*   
## IMDATE 5.049e-06 1.275e-06 3.961 7.48e-05 \*\*\*  
## IDISTM -6.264e-02 1.262e-02 -4.965 6.92e-07 \*\*\*  
## ISEASNBOS NA NA NA NA   
## ISEASNCAL 1.641e-01 1.636e-01 1.004 0.315626   
## ISEASNCOM NA NA NA NA   
## ISEASNEAS -6.556e-02 1.371e-01 -0.478 0.632588   
## ISEASNEDY -1.141e-01 1.052e+00 -0.108 0.913608   
## ISEASNFAD -8.929e-02 1.480e-01 -0.603 0.546204   
## ISEASNGRD -2.846e-03 1.502e-01 -0.019 0.984886   
## ISEASNHAL -1.028e-01 1.911e-01 -0.538 0.590540   
## ISEASNHNK 5.538e-01 1.737e-01 3.188 0.001436 \*\*   
## ISEASNJNY -3.201e-02 2.376e-01 -0.135 0.892813   
## ISEASNMOD -9.383e-02 1.142e-01 -0.821 0.411378   
## ISEASNNYD 8.577e-02 6.949e-01 0.123 0.901772   
## ISEASNPAS -1.987e-01 5.697e-01 -0.349 0.727297   
## ISEASNSEC NA NA NA NA   
## ISEASNSTP NA NA NA NA   
## ISEASNTHG 2.415e-01 9.956e-01 0.243 0.808312   
## ISEASNVAL -3.849e-02 1.190e-01 -0.323 0.746370   
## ISEASNXMS 6.255e-01 7.902e-02 7.915 2.60e-15 \*\*\*  
## INLRTL 2.292e-02 1.317e-02 1.740 0.081886 .   
## IATRB3AU 4.650e-01 3.294e-01 1.412 0.158106   
## IATRB3CA 5.582e-03 2.010e-01 0.028 0.977838   
## IATRB3CH -1.855e-01 1.634e-01 -1.135 0.256363   
## IATRB3CN 3.269e-02 3.813e-02 0.857 0.391212   
## IATRB3DE 4.172e-01 3.308e-01 1.261 0.207359   
## IATRB3EN -5.828e-03 7.004e-01 -0.008 0.993361   
## IATRB3ES 1.184e-02 1.718e-01 0.069 0.945043   
## IATRB3FR 3.328e-02 1.222e-01 0.272 0.785374   
## IATRB3GB -8.212e-02 6.564e-02 -1.251 0.210934   
## IATRB3HK -7.937e-02 3.168e-01 -0.251 0.802200   
## IATRB3HU 3.541e-02 2.306e-01 0.154 0.877948   
## IATRB3ID 5.398e-01 2.364e-01 2.283 0.022431 \*   
## IATRB3IL -1.392e-01 1.416e-01 -0.983 0.325780   
## IATRB3IN 5.488e-02 6.713e-02 0.818 0.413594   
## IATRB3IT 9.203e-02 7.542e-02 1.220 0.222394   
## IATRB3JP -3.212e-01 1.683e-01 -1.908 0.056383 .   
## IATRB3KR 1.266e-01 7.738e-02 1.637 0.101744   
## IATRB3LB -3.293e-02 1.961e-01 -0.168 0.866662   
## IATRB3MA 5.788e-02 5.725e-01 0.101 0.919471   
## IATRB3MG 8.600e-02 2.349e-01 0.366 0.714242   
## IATRB3MX 1.904e-04 1.400e-01 0.001 0.998915   
## IATRB3MY 1.816e-01 3.041e-01 0.597 0.550545   
## IATRB3NP -3.023e-02 2.859e-01 -0.106 0.915791   
## IATRB3PE -9.357e-02 2.781e-01 -0.336 0.736498   
## IATRB3PH 1.618e-01 1.367e-01 1.184 0.236404   
## IATRB3PT -4.311e-01 5.743e-01 -0.751 0.452880   
## IATRB3SG 1.559e-01 5.697e-01 0.274 0.784306   
## IATRB3SL 2.932e+00 6.984e-01 4.199 2.70e-05 \*\*\*  
## IATRB3TH -1.052e-01 6.878e-02 -1.529 0.126172   
## IATRB3TR -5.800e-02 1.464e-01 -0.396 0.692027   
## IATRB3TW 4.531e-01 8.542e-02 5.304 1.14e-07 \*\*\*  
## IATRB3US 2.400e-02 3.783e-02 0.634 0.525794   
## IATRB3VN -1.046e-01 1.661e-01 -0.630 0.528623   
## IPRCCHN -1.712e+00 2.125e-01 -8.057 8.29e-16 \*\*\*  
## IPRCCHR -8.131e-02 1.338e-01 -0.608 0.543351   
## IPRCSTN 1.513e-01 6.972e-02 2.171 0.029976 \*   
## IPRCSTP 8.197e-02 6.911e-02 1.186 0.235598   
## IPRCSTR 1.276e-01 7.486e-02 1.704 0.088316 .   
## ICORGPCARDS 7.772e-02 3.809e-01 0.204 0.838312   
## ICORGPCOLLE -7.233e-01 1.453e+00 -0.498 0.618662   
## ICORGPGIFTS 9.341e-02 1.261e+00 0.074 0.940972   
## ICORGPHOMDE -2.411e-01 1.139e+00 -0.212 0.832316   
## ICORGPMEMRY 1.314e-01 9.823e-01 0.134 0.893592   
## ICORGPSEASN 1.981e+01 1.526e+00 12.979 < 2e-16 \*\*\*  
## ICORGPSTATN 1.964e-01 4.173e-01 0.471 0.637816   
## ICORGPWRAP NA NA NA NA   
## ILEAD -4.087e-04 6.257e-04 -0.653 0.513651   
## IMCRDT -2.062e-06 1.040e-06 -1.982 0.047504 \*   
## w\_babi -6.471e-02 6.995e-02 -0.925 0.354894   
## w\_bag 6.709e-02 5.957e-02 1.126 0.260083   
## w\_bdi -6.140e-02 9.167e-02 -0.670 0.503003   
## w\_bird -2.516e-02 6.774e-02 -0.371 0.710340   
## w\_birthday -3.217e-02 5.486e-02 -0.586 0.557576   
## w\_black 1.042e-01 6.019e-02 1.730 0.083567 .   
## w\_blue 3.101e-02 6.566e-02 0.472 0.636720   
## w\_box 1.001e-01 6.956e-02 1.440 0.149974   
## w\_bracelet 7.659e-03 8.733e-02 0.088 0.930115   
## w\_butterfli 7.875e-03 7.566e-02 0.104 0.917107   
## w\_cake 4.926e-03 7.765e-02 0.063 0.949424   
## w\_candl 2.488e-02 6.959e-02 0.357 0.720750   
## w\_card -2.313e-02 5.645e-02 -0.410 0.682034   
## w\_cardx -8.375e-02 1.364e-01 -0.614 0.539171   
## w\_conv -2.014e-02 1.155e-01 -0.174 0.861621   
## w\_crystal 3.733e-02 6.818e-02 0.548 0.583975   
## w\_dog -2.931e-02 7.964e-02 -0.368 0.712859   
## w\_dress -5.527e-02 7.617e-02 -0.726 0.468102   
## w\_earring 2.240e-02 8.482e-02 0.264 0.791737   
## w\_floral 4.204e-02 5.858e-02 0.718 0.472988   
## w\_flower 1.310e-02 4.506e-02 0.291 0.771240   
## w\_gem -2.164e-03 7.207e-02 -0.030 0.976052   
## w\_general 1.015e-01 8.946e-02 1.135 0.256444   
## w\_girl 6.299e-02 6.294e-02 1.001 0.316995   
## w\_glass 3.033e-03 7.231e-02 0.042 0.966541   
## w\_glitter -1.137e-01 8.111e-02 -1.401 0.161102   
## w\_gold -2.375e-02 5.783e-02 -0.411 0.681270   
## w\_handmad 1.411e-02 8.271e-02 0.171 0.864509   
## w\_happi 8.896e-03 6.581e-02 0.135 0.892476   
## w\_heart 4.245e-02 6.333e-02 0.670 0.502715   
## w\_love 2.756e-02 5.791e-02 0.476 0.634192   
## w\_med -9.889e-02 8.497e-02 -1.164 0.244547   
## w\_mini -3.900e-02 8.759e-02 -0.445 0.656118   
## w\_mom 4.074e-02 8.353e-02 0.488 0.625716   
## w\_neck -2.094e-02 8.210e-02 -0.255 0.798685   
## w\_necklac -2.575e-02 8.588e-02 -0.300 0.764273   
## w\_note 4.264e-02 6.388e-02 0.667 0.504498   
## w\_pink -9.867e-03 6.913e-02 -0.143 0.886505   
## w\_print -6.893e-02 9.067e-02 -0.760 0.447112   
## w\_red 1.124e-01 6.708e-02 1.676 0.093841 .   
## w\_ribbon 1.827e-01 9.550e-02 1.913 0.055766 .   
## w\_set 4.321e-02 6.610e-02 0.654 0.513294   
## w\_silver 1.852e-01 7.444e-02 2.488 0.012850 \*   
## w\_soap 3.945e-02 7.471e-02 0.528 0.597451   
## w\_thank -4.305e-02 8.394e-02 -0.513 0.608027   
## w\_tree 1.490e-01 6.718e-02 2.217 0.026629 \*   
## w\_vintag -1.043e-01 8.175e-02 -1.276 0.202101   
## w\_wed -5.905e-02 8.070e-02 -0.732 0.464351   
## w\_white 8.121e-02 7.446e-02 1.091 0.275458   
## w\_xbc -6.027e-02 9.314e-02 -0.647 0.517611   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9813 on 18223 degrees of freedom  
## Multiple R-squared: 0.4586, Adjusted R-squared: 0.4506   
## F-statistic: 57.38 on 269 and 18223 DF, p-value: < 2.2e-16

Remove factor levels not correlated:

Inv\_final$IMFGNO[!(Inv\_final$IMFGNO %in% c("4772", "4773", "4777", "9361", "9739"))]<-""  
Inv\_final$IFINLN[!(Inv\_final$IFINLN %in% c("OTHER"))]<-""  
Inv\_final$ISTYPE[!(Inv\_final$ISTYPE %in% c("05", "GW"))]<-""  
Inv\_final$ISEASN[!(Inv\_final$ISEASN %in% c("HNK", "XMS"))]<-""  
Inv\_final$IATRB3[!(Inv\_final$IATRB3 %in% c("ID", "JP", "SL", "TW"))]<-""

Minimal Linear Regression model (all non-empty attributes, removing non-correlated and colinear attributes, and non-correlated words) Attributes removed: ITKTTR, IWGHT, IMDATE, ILEAD, IMXSTK, IRPLCD, IPRCCH, IPRCST, ICORGP. Words kept: Black, Red, Ribbon, Silver, Tree R-squared: 0.3724

##   
## Call:  
## lm(formula = IBWK01\_sum ~ IBHAND\_sum + IBWK02\_sum + IBWK03\_sum +   
## IBWK04\_sum + IBWK05\_sum + IBWK06\_sum + IBWK07\_sum + IBWK08\_sum +   
## ASNUM + IDEPT + ISDEPT + ICLAS + ISCLAS + IMFGNO + BYRNUM +   
## ITKTN + ILBLTR + IFINLN + IMINPK + ISTDPK + IDSPLY + IVPLHI +   
## IDISTM + ISEASN + INLRTL + IATRB3 + IMCRDT + w\_black + w\_red +   
## w\_ribbon + w\_silver + w\_tree, data = Inv\_final)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.564 -0.110 0.003 0.087 79.920   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.322e-02 9.671e-02 0.240 0.810231   
## IBHAND\_sum -1.147e-02 4.407e-04 -26.025 < 2e-16 \*\*\*  
## IBWK02\_sum 3.251e-01 1.298e-02 25.043 < 2e-16 \*\*\*  
## IBWK03\_sum 4.540e-01 1.653e-02 27.462 < 2e-16 \*\*\*  
## IBWK04\_sum 3.441e-01 1.197e-02 28.747 < 2e-16 \*\*\*  
## IBWK05\_sum 7.247e-02 1.612e-02 4.496 6.97e-06 \*\*\*  
## IBWK06\_sum 3.244e-01 1.683e-02 19.275 < 2e-16 \*\*\*  
## IBWK07\_sum 3.275e-01 2.020e-02 16.210 < 2e-16 \*\*\*  
## IBWK08\_sum 5.682e-02 1.525e-02 3.727 0.000194 \*\*\*  
## ASNUM -4.870e-08 9.771e-08 -0.498 0.618170   
## IDEPT 9.028e-04 1.511e-04 5.976 2.33e-09 \*\*\*  
## ISDEPT -1.246e-03 6.026e-04 -2.069 0.038599 \*   
## ICLAS -5.342e-04 2.335e-04 -2.288 0.022157 \*   
## ISCLAS -1.102e-03 4.001e-04 -2.754 0.005894 \*\*   
## IMFGNO4772 -7.249e-01 1.741e-01 -4.165 3.13e-05 \*\*\*  
## IMFGNO4773 -6.929e-01 1.417e-01 -4.890 1.02e-06 \*\*\*  
## IMFGNO4777 -4.204e-01 9.600e-02 -4.379 1.20e-05 \*\*\*  
## IMFGNO9361 -3.586e-01 1.548e-01 -2.317 0.020540 \*   
## IMFGNO9739 2.663e-01 8.578e-02 3.105 0.001906 \*\*   
## BYRNUM -1.166e-02 5.339e-03 -2.184 0.028978 \*   
## ITKTN 2.968e-01 2.086e-01 1.423 0.154771   
## ILBLTR -3.264e-01 2.075e-01 -1.573 0.115752   
## IFINLNOTHER 7.285e+00 3.561e-01 20.455 < 2e-16 \*\*\*  
## IMINPK 9.241e-03 1.656e-03 5.580 2.44e-08 \*\*\*  
## ISTDPK -1.335e-04 6.834e-05 -1.954 0.050771 .   
## IDSPLY 6.225e-03 2.317e-03 2.686 0.007233 \*\*   
## IVPLHI -3.495e-02 8.198e-03 -4.263 2.03e-05 \*\*\*  
## IDISTM -3.864e-02 8.216e-03 -4.703 2.58e-06 \*\*\*  
## ISEASNHNK 9.457e-01 1.310e-01 7.222 5.33e-13 \*\*\*  
## ISEASNXMS 7.157e-01 4.401e-02 16.263 < 2e-16 \*\*\*  
## INLRTL 2.320e-02 1.054e-02 2.202 0.027664 \*   
## IATRB3ID 6.113e-01 2.355e-01 2.596 0.009449 \*\*   
## IATRB3JP -3.254e-01 1.600e-01 -2.034 0.041918 \*   
## IATRB3SL 3.023e+00 7.429e-01 4.070 4.73e-05 \*\*\*  
## IATRB3TW 4.003e-01 7.801e-02 5.131 2.91e-07 \*\*\*  
## IMCRDT -2.672e-07 7.041e-07 -0.379 0.704373   
## w\_black 1.126e-01 6.378e-02 1.765 0.077579 .   
## w\_red 1.376e-01 7.097e-02 1.939 0.052538 .   
## w\_ribbon 1.216e-01 9.928e-02 1.225 0.220623   
## w\_silver 2.049e-01 7.880e-02 2.600 0.009329 \*\*   
## w\_tree 1.731e-01 7.136e-02 2.426 0.015282 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.05 on 18452 degrees of freedom  
## Multiple R-squared: 0.3724, Adjusted R-squared: 0.371   
## F-statistic: 273.7 on 40 and 18452 DF, p-value: < 2.2e-16

# Prepare data for model evaluation

Break data set up into Training and Test sets (70-30).

set.seed(123)  
index<-sample(1:nrow(Inv\_final),size=0.7\*nrow(Inv\_final))  
  
Inv\_final\_train<-Inv\_final[index,]  
Inv\_final\_test<-Inv\_final[-index,]

Remove factor levels from test set not in training set.

# Create models and evaluate

Calculate Baseline Model

best.guess<-mean(Inv\_final\_train$IBWK01\_sum)  
RMSE.baseline<-sqrt(mean((best.guess-Inv\_final\_test$IBWK01\_sum)^2))  
message('RMSE: ', RMSE.baseline)

## RMSE: 1.9548638714503

MAE.baseline<-mean(abs(best.guess-Inv\_final\_test$IBWK01\_sum))  
message('MAE: ', MAE.baseline)

## MAE: 0.433979110779536

Create minimal Linear Regression model on Training Data

##   
## Call:  
## lm(formula = IBWK01\_sum ~ IBHAND\_sum + IBWK02\_sum + IBWK03\_sum +   
## IBWK04\_sum + IBWK05\_sum + IBWK06\_sum + IBWK07\_sum + IBWK08\_sum +   
## ASNUM + IDEPT + ISDEPT + ICLAS + ISCLAS + IMFGNO + BYRNUM +   
## ITKTN + ILBLTR + IFINLN + IMINPK + ISTDPK + IDSPLY + IVPLHI +   
## IDISTM + ISEASN + INLRTL + IATRB3 + IMCRDT + w\_black + w\_red +   
## w\_ribbon + w\_silver + w\_tree, data = Inv\_final\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.5768 -0.1247 -0.0383 0.0051 15.9234   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.436e-01 8.528e-02 -5.202 2.01e-07 \*\*\*  
## IBHAND\_sum 1.563e-03 4.049e-04 3.862 0.000113 \*\*\*  
## IBWK02\_sum 3.154e-01 1.096e-02 28.770 < 2e-16 \*\*\*  
## IBWK03\_sum 3.130e-01 1.412e-02 22.170 < 2e-16 \*\*\*  
## IBWK04\_sum 6.317e-02 1.010e-02 6.254 4.14e-10 \*\*\*  
## IBWK05\_sum 1.100e-01 1.415e-02 7.775 8.11e-15 \*\*\*  
## IBWK06\_sum 1.503e-01 1.433e-02 10.489 < 2e-16 \*\*\*  
## IBWK07\_sum 6.628e-02 1.788e-02 3.708 0.000210 \*\*\*  
## IBWK08\_sum 5.956e-03 1.240e-02 0.480 0.631092   
## ASNUM -1.550e-07 8.573e-08 -1.808 0.070596 .   
## IDEPT 4.960e-04 1.322e-04 3.751 0.000177 \*\*\*  
## ISDEPT -6.830e-04 5.275e-04 -1.295 0.195393   
## ICLAS -2.583e-04 2.048e-04 -1.262 0.207074   
## ISCLAS -1.100e-04 3.487e-04 -0.316 0.752346   
## IMFGNO4772 -7.176e-01 1.489e-01 -4.821 1.44e-06 \*\*\*  
## IMFGNO4773 -6.749e-01 1.183e-01 -5.705 1.19e-08 \*\*\*  
## IMFGNO4777 -3.941e-01 8.723e-02 -4.518 6.31e-06 \*\*\*  
## IMFGNO9361 -3.267e-01 1.399e-01 -2.336 0.019514 \*   
## IMFGNO9739 4.846e-01 7.691e-02 6.300 3.07e-10 \*\*\*  
## BYRNUM -8.944e-03 4.702e-03 -1.902 0.057169 .   
## ITKTN 2.424e-01 1.900e-01 1.276 0.202007   
## ILBLTR -2.435e-01 1.889e-01 -1.289 0.197427   
## IFINLNOTHER 3.970e-01 3.045e-01 1.304 0.192358   
## IMINPK 8.514e-03 2.177e-03 3.910 9.26e-05 \*\*\*  
## ISTDPK -1.139e-04 5.908e-05 -1.928 0.053876 .   
## IDSPLY 2.197e-04 2.049e-03 0.107 0.914604   
## IVPLHI -1.759e-02 7.388e-03 -2.380 0.017305 \*   
## IDISTM -4.255e-03 7.238e-03 -0.588 0.556684   
## ISEASNHNK 4.847e-01 1.144e-01 4.235 2.30e-05 \*\*\*  
## ISEASNXMS 7.054e-01 3.883e-02 18.165 < 2e-16 \*\*\*  
## INLRTL 1.159e-03 9.336e-03 0.124 0.901214   
## IATRB3ID 8.128e-01 2.002e-01 4.059 4.95e-05 \*\*\*  
## IATRB3JP 3.760e-01 1.388e-01 2.709 0.006765 \*\*   
## IATRB3SL 2.843e+00 5.468e-01 5.199 2.03e-07 \*\*\*  
## IATRB3TW 2.000e-01 6.809e-02 2.938 0.003310 \*\*   
## IMCRDT 3.240e-06 6.202e-07 5.224 1.78e-07 \*\*\*  
## w\_black 9.834e-02 5.549e-02 1.772 0.076381 .   
## w\_red 2.334e-01 6.190e-02 3.771 0.000163 \*\*\*  
## w\_ribbon 2.371e-01 8.778e-02 2.700 0.006933 \*\*   
## w\_silver 2.059e-01 7.086e-02 2.906 0.003670 \*\*   
## w\_tree 1.314e-01 6.343e-02 2.072 0.038277 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7725 on 12904 degrees of freedom  
## Multiple R-squared: 0.313, Adjusted R-squared: 0.3109   
## F-statistic: 147 on 40 and 12904 DF, p-value: < 2.2e-16

Calculate errors on LR model

## RMSE: 1.76019538031604

## MAE: 0.352834516139955

Create Decision Tree model

## Call:  
## rpart(formula = IBWK01\_sum ~ IBHAND\_sum + IBWK02\_sum + IBWK03\_sum +   
## IBWK04\_sum + IBWK05\_sum + IBWK06\_sum + IBWK07\_sum + IBWK08\_sum +   
## ASNUM + IDEPT + ISDEPT + ICLAS + ISCLAS + IMFGNO + BYRNUM +   
## ITKTN + ILBLTR + IFINLN + IMINPK + ISTDPK + IDSPLY + IVPLHI +   
## IDISTM + ISEASN + INLRTL + IATRB3 + IMCRDT + w\_black + w\_red +   
## w\_ribbon + w\_silver + w\_tree, data = Inv\_final\_train)  
## n= 12945   
##   
## CP nsplit rel error xerror xstd  
## 1 0.10488562 0 1.0000000 1.0000851 0.08820635  
## 2 0.05076975 1 0.8951144 0.8971839 0.07896143  
## 3 0.04133386 2 0.8443446 0.9182936 0.07919399  
## 4 0.01569497 3 0.8030108 0.8445828 0.07295950  
## 5 0.01161530 5 0.7716208 0.8527025 0.07281953  
## 6 0.01107939 6 0.7600055 0.8478713 0.07285797  
## 7 0.01045493 7 0.7489261 0.8467747 0.07219435  
## 8 0.01000000 8 0.7384712 0.8470500 0.07219315  
##   
## Variable importance  
## IBWK02\_sum ISEASN IBWK03\_sum ISDEPT IBWK04\_sum IMFGNO   
## 44 15 14 7 6 4   
## IBHAND\_sum IBWK07\_sum IDISTM IDSPLY ISCLAS IBWK06\_sum   
## 3 1 1 1 1 1   
## IBWK05\_sum   
## 1   
##   
## Node number 1: 12945 observations, complexity param=0.1048856  
## mean=0.2387022, MSE=0.8660031   
## left son=2 (12503 obs) right son=3 (442 obs)  
## Primary splits:  
## IBWK02\_sum < 1.5 to the left, improve=0.10488560, (0 missing)  
## IBWK03\_sum < 7.5 to the left, improve=0.08332687, (0 missing)  
## ISEASN splits as LRR, improve=0.07283051, (0 missing)  
## IBHAND\_sum < -36.5 to the right, improve=0.04495054, (0 missing)  
## IBWK04\_sum < 0.5 to the left, improve=0.04457868, (0 missing)  
## Surrogate splits:  
## IBWK03\_sum < 4.5 to the left, agree=0.967, adj=0.020, (0 split)  
## IBWK06\_sum < 3.5 to the left, agree=0.966, adj=0.011, (0 split)  
## IBHAND\_sum < -56.5 to the right, agree=0.966, adj=0.009, (0 split)  
## IMCRDT < 30209 to the right, agree=0.966, adj=0.009, (0 split)  
## IBWK05\_sum < 9.5 to the left, agree=0.966, adj=0.005, (0 split)  
##   
## Node number 2: 12503 observations, complexity param=0.04133386  
## mean=0.1820363, MSE=0.5438443   
## left son=4 (11482 obs) right son=5 (1021 obs)  
## Primary splits:  
## ISEASN splits as LRR, improve=0.06814574, (0 missing)  
## IBWK03\_sum < 0.5 to the left, improve=0.04558609, (0 missing)  
## ISDEPT < 70 to the left, improve=0.04179843, (0 missing)  
## IBWK02\_sum < 0.5 to the left, improve=0.04021285, (0 missing)  
## IBWK04\_sum < 0.5 to the left, improve=0.03483760, (0 missing)  
## Surrogate splits:  
## ISDEPT < 70 to the left, agree=0.950, adj=0.382, (0 split)  
## IMFGNO splits as LRRRRR, agree=0.942, adj=0.292, (0 split)  
## w\_tree < 0.5 to the left, agree=0.921, adj=0.030, (0 split)  
## IMCRDT < 151107 to the left, agree=0.919, adj=0.014, (0 split)  
## ASNUM < 605367.5 to the left, agree=0.919, adj=0.003, (0 split)  
##   
## Node number 3: 442 observations, complexity param=0.05076975  
## mean=1.841629, MSE=7.31881   
## left son=6 (432 obs) right son=7 (10 obs)  
## Primary splits:  
## IBWK02\_sum < 8.5 to the left, improve=0.17593970, (0 missing)  
## IBHAND\_sum < -37 to the right, improve=0.13627340, (0 missing)  
## IBWK03\_sum < 4.5 to the left, improve=0.13468950, (0 missing)  
## IBWK05\_sum < 4.5 to the left, improve=0.08565021, (0 missing)  
## IBWK06\_sum < 2.5 to the left, improve=0.06734282, (0 missing)  
## Surrogate splits:  
## IBHAND\_sum < -205.5 to the right, agree=0.982, adj=0.2, (0 split)  
## IBWK03\_sum < 16 to the left, agree=0.982, adj=0.2, (0 split)  
## IBWK04\_sum < 6.5 to the left, agree=0.980, adj=0.1, (0 split)  
## IBWK07\_sum < 5.5 to the left, agree=0.980, adj=0.1, (0 split)  
##   
## Node number 4: 11482 observations, complexity param=0.01569497  
## mean=0.1246299, MSE=0.3524347   
## left son=8 (10724 obs) right son=9 (758 obs)  
## Primary splits:  
## IBWK04\_sum < 0.5 to the left, improve=0.04095379, (0 missing)  
## IBWK03\_sum < 0.5 to the left, improve=0.03962678, (0 missing)  
## IBWK02\_sum < 0.5 to the left, improve=0.03073654, (0 missing)  
## IBWK05\_sum < 0.5 to the left, improve=0.03068061, (0 missing)  
## IBWK06\_sum < 0.5 to the left, improve=0.01993223, (0 missing)  
## Surrogate splits:  
## IBHAND\_sum < -19.5 to the right, agree=0.934, adj=0.003, (0 split)  
## IBWK03\_sum < 4.5 to the left, agree=0.934, adj=0.003, (0 split)  
## IBWK05\_sum < 4.5 to the left, agree=0.934, adj=0.001, (0 split)  
## IBWK06\_sum < 5.5 to the left, agree=0.934, adj=0.001, (0 split)  
##   
## Node number 5: 1021 observations, complexity param=0.01107939  
## mean=0.82762, MSE=2.242567   
## left son=10 (901 obs) right son=11 (120 obs)  
## Primary splits:  
## IBWK03\_sum < 0.5 to the left, improve=0.05424582, (0 missing)  
## IBWK05\_sum < 0.5 to the left, improve=0.03953062, (0 missing)  
## ASNUM < 13925.5 to the left, improve=0.03741260, (0 missing)  
## IBWK02\_sum < 0.5 to the left, improve=0.03712765, (0 missing)  
## ISDEPT < 6 to the left, improve=0.03492912, (0 missing)  
## Surrogate splits:  
## IBWK08\_sum < 1.5 to the left, agree=0.884, adj=0.017, (0 split)  
##   
## Node number 6: 432 observations, complexity param=0.0116153  
## mean=1.668981, MSE=5.068667   
## left son=12 (287 obs) right son=13 (145 obs)  
## Primary splits:  
## ISEASN splits as LLR, improve=0.05946678, (0 missing)  
## IBWK03\_sum < 3.5 to the left, improve=0.05921541, (0 missing)  
## IDISTM < 2.5 to the right, improve=0.05470957, (0 missing)  
## IDSPLY < 0.5 to the right, improve=0.04857995, (0 missing)  
## ISDEPT < 11 to the left, improve=0.04848684, (0 missing)  
## Surrogate splits:  
## ISDEPT < 25 to the left, agree=0.873, adj=0.621, (0 split)  
## IDSPLY < 0.5 to the right, agree=0.759, adj=0.283, (0 split)  
## IDISTM < 1 to the right, agree=0.759, adj=0.283, (0 split)  
## IMFGNO splits as L-RRRR, agree=0.748, adj=0.248, (0 split)  
## ISCLAS < 1.5 to the right, agree=0.741, adj=0.228, (0 split)  
##   
## Node number 7: 10 observations  
## mean=9.3, MSE=47.61   
##   
## Node number 8: 10724 observations  
## mean=0.0926893, MSE=0.2321771   
##   
## Node number 9: 758 observations, complexity param=0.01569497  
## mean=0.5765172, MSE=1.835174   
## left son=18 (750 obs) right son=19 (8 obs)  
## Primary splits:  
## IBWK03\_sum < 3.5 to the left, improve=0.13383180, (0 missing)  
## IBHAND\_sum < -18 to the right, improve=0.05466309, (0 missing)  
## ISCLAS < 121.5 to the left, improve=0.04583173, (0 missing)  
## IBWK04\_sum < 1.5 to the left, improve=0.03256573, (0 missing)  
## IBWK06\_sum < 1.5 to the left, improve=0.02362210, (0 missing)  
##   
## Node number 10: 901 observations  
## mean=0.700333, MSE=1.916859   
##   
## Node number 11: 120 observations  
## mean=1.783333, MSE=3.653056   
##   
## Node number 12: 287 observations, complexity param=0.01045493  
## mean=1.278746, MSE=3.302092   
## left son=24 (272 obs) right son=25 (15 obs)  
## Primary splits:  
## IBWK03\_sum < 3.5 to the left, improve=0.12367200, (0 missing)  
## IBWK05\_sum < 0.5 to the left, improve=0.09793987, (0 missing)  
## IBHAND\_sum < 113.5 to the left, improve=0.08935954, (0 missing)  
## IBWK07\_sum < 1.5 to the left, improve=0.07985649, (0 missing)  
## IBWK06\_sum < 2.5 to the left, improve=0.07222686, (0 missing)  
## Surrogate splits:  
## IBWK04\_sum < 6.5 to the left, agree=0.955, adj=0.133, (0 split)  
## IBWK05\_sum < 9 to the left, agree=0.955, adj=0.133, (0 split)  
## IBWK06\_sum < 7 to the left, agree=0.951, adj=0.067, (0 split)  
##   
## Node number 13: 145 observations  
## mean=2.441379, MSE=7.667253   
##   
## Node number 18: 750 observations  
## mean=0.5253333, MSE=1.134692   
##   
## Node number 19: 8 observations  
## mean=5.375, MSE=44.23438   
##   
## Node number 24: 272 observations  
## mean=1.128676, MSE=2.663589   
##   
## Node number 25: 15 observations  
## mean=4, MSE=7.066667

Calculate errors on DT model

## RMSE: 1.78428809643065

## MAE: 0.343415309186539

Prune resulting model

##   
## Regression tree:  
## rpart(formula = IBWK01\_sum ~ IBHAND\_sum + IBWK02\_sum + IBWK03\_sum +   
## IBWK04\_sum + IBWK05\_sum + IBWK06\_sum + IBWK07\_sum + IBWK08\_sum +   
## ASNUM + IDEPT + ISDEPT + ICLAS + ISCLAS + IMFGNO + BYRNUM +   
## ITKTN + ILBLTR + IFINLN + IMINPK + ISTDPK + IDSPLY + IVPLHI +   
## IDISTM + ISEASN + INLRTL + IATRB3 + IMCRDT + w\_black + w\_red +   
## w\_ribbon + w\_silver + w\_tree, data = Inv\_final\_train)  
##   
## Variables actually used in tree construction:  
## [1] IBWK02\_sum IBWK03\_sum IBWK04\_sum ISEASN   
##   
## Root node error: 11210/12945 = 0.866  
##   
## n= 12945   
##   
## CP nsplit rel error xerror xstd  
## 1 0.104886 0 1.00000 1.00009 0.088206  
## 2 0.050770 1 0.89511 0.89718 0.078961  
## 3 0.041334 2 0.84434 0.91829 0.079194  
## 4 0.015695 3 0.80301 0.84458 0.072960  
## 5 0.011615 5 0.77162 0.85270 0.072820  
## 6 0.011079 6 0.76001 0.84787 0.072858  
## 7 0.010455 7 0.74893 0.84677 0.072194  
## 8 0.010000 8 0.73847 0.84705 0.072193

Calculate errors on pruned DT model

## RMSE: 1.79187762404969

## MAE: 0.362075737970264

Create Random Forest model

model\_rf=randomForest(IBWK01\_sum~IBHAND\_sum+IBWK02\_sum+IBWK03\_sum+IBWK04\_sum+IBWK05\_sum+IBWK06\_sum+IBWK07\_sum+IBWK08\_sum+ASNUM+IDEPT+ISDEPT+ICLAS+ISCLAS+IMFGNO+BYRNUM+ITKTN+ILBLTR+IFINLN+IMINPK+ISTDPK+IDSPLY+IVPLHI+IDISTM+ISEASN+INLRTL+IATRB3+IMCRDT+w\_black+w\_red+w\_ribbon+w\_silver+w\_tree, data=Inv\_final\_train,importance=TRUE, ntree=1000)  
#summary(model\_rf)  
  
message('Minimum MSE tree count:', which.min(model\_rf$mse))

## Minimum MSE tree count:854

imp <- as.data.frame(sort(importance(model\_rf)[,1],decreasing = TRUE),optional = T)  
names(imp) <- "% Inc MSE"  
message('Importance of independent variables:')

## Importance of independent variables:

imp

## % Inc MSE  
## ISEASN 49.5671322  
## IBHAND\_sum 47.0998715  
## IMCRDT 42.5984003  
## ISTDPK 27.9817961  
## ICLAS 26.3686505  
## IBWK02\_sum 26.2101749  
## IBWK03\_sum 22.5720300  
## IMINPK 21.2217396  
## ISDEPT 20.4357009  
## INLRTL 20.2323786  
## IDEPT 19.6573954  
## ISCLAS 18.8007750  
## ASNUM 17.6694103  
## IBWK07\_sum 16.8975064  
## IBWK04\_sum 15.7505443  
## IDISTM 15.2256035  
## IATRB3 14.2222129  
## ITKTN 13.6860436  
## BYRNUM 13.0318743  
## IMFGNO 12.5187499  
## IDSPLY 12.2581882  
## ILBLTR 12.2106645  
## IVPLHI 8.6764683  
## IBWK05\_sum 8.2892117  
## IBWK08\_sum 7.9780750  
## IBWK06\_sum 7.5627022  
## w\_black 4.6331126  
## w\_ribbon 3.2279347  
## IFINLN 0.3502756  
## w\_silver -5.0610544  
## w\_red -8.0433318  
## w\_tree -11.3833657

Calculate errors on RF model

## RMSE: 1.7513850799721

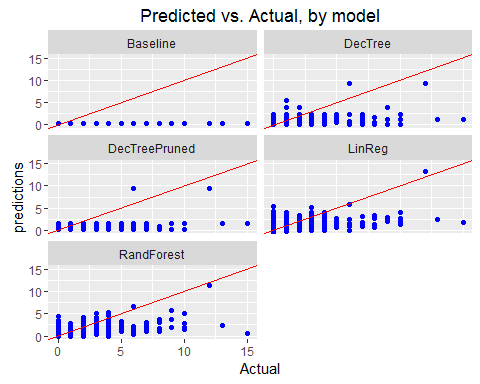
## MAE: 0.302763730772943

# Putting it all together!

## Method RMSE MAE  
## 1 Baseline 1.95 0.43  
## 2 Linear Regression 1.76 0.35  
## 3 Full tree 1.78 0.34  
## 4 Pruned tree 1.79 0.36  
## 5 Random forest 1.75 0.30

## First few predictions:

## Actual Baseline LinReg DecTree DecTreePruned RandForest  
## 2 2 0.24 2.97 0.09 0.12 1.34  
## 3 0 0.24 3.00 0.09 0.12 1.39  
## 5 0 0.24 -0.08 0.09 0.12 0.08  
## 7 0 0.24 0.67 0.70 0.83 1.60  
## 13 1 0.24 0.00 0.53 0.12 0.37  
## 15 2 0.24 0.01 0.09 0.12 0.44



## Sales Units range is cut off at 15

# Summary

From RMSE and MAE measurements, it appears that the Random Forest model is the best, which is to be expected. Looking at the graphs of Predicted vs. Actual, Random Forest looks to push closer to ideal. More input is needed to get a better prediction of R-squared of 0.37.

However, before removing all of the non-correlated attributes and description words, the R-squared was 0.46, and the best RMSE was higher (~1.80) than this minimal model.

It is difficult guaging which of the independent variables are the most important, as the Linear Regression lowest t valued variables don't correspond to the lowest Random Forest MSE valued variables. Perhaps developing the RF model on the full training set and then comparing to the LR model developed on the full training set. That would give a better representation of the full story.