ettain group

Shyni

3/4/2020

The attached data set contains inventory accuracy records from a number of Grocery stores. The ‘adjustment unit quantity’ is the unit variance that 3rd party auditors confirmed upon visiting a store, e.g., the first example below (-1) means that the auditing company found one less bowl of tropical fruit than the stores perpetual inventory had on record. The subsequent example (+14) means that the auditors counted fourteen more peach fruit bowls than the stores perpetual inventory. Your objective is to work the data set in R: i) 3 ‘actionable’ insights on inventory accuracy (by store, commodity, variance amounts, etc..) ii) Plots with supporting insights that visualize your story in ggplot Set Library

```
library(readr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────── tidyverse 1.2.1 ──
```

```
## ✓ ggplot2 3.2.0     ✓ purrr   0.3.3
## ✓ tibble  2.1.3     ✓ dplyr   0.8.3
## ✓ tidyr   1.0.0     ✓ stringr 1.4.0
## ✓ ggplot2 3.2.0     ✓ forcats 0.3.0
```

```
## ── Conflicts ────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): un
able to load shared object '/Library/Frameworks/R.framework/Resource
s/modules//R_X11.so':
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11
.so, 6): Library not loaded: /opt/X11/lib/libSM.6.dylib
##   Referenced from: /Library/Frameworks/R.framework/Resources/modu
les//R_X11.so
##   Reason: image not found
```

```
## Could not load tcltk.  Will use slower R code instead.
```

```
## Loading required package: RSQLite
```

```
library(readxl)
library(ggplot2)
```

Import Data

```
data_xlsb <- read_excel("~/Desktop/InventoryData.xlsx")
data1 <- data_xlsb
```

Data Wrangling

```
colnames(data1)[colnames(data1)=="Store #"] <- "Store_number"
colnames(data1)[colnames(data1)=="Commodity Name" ] <- "Commodity_Na
me"
colnames(data1)[colnames(data1)=="Base GTIN Number"] <- "Base_GTIN_N
umber"
colnames(data1)[colnames(data1)=="Base GTIN Description"] <- "Base_G
TIN_Description"
colnames(data1)[colnames(data1)=="Adjustment Unit Quantity"] <- "Adj
ustment_Unit_Quantity"
```

Getting the metrics about data types, zeros, infinite numbers, and missing values: Get A Summary For The Given Data Frame (O Vector). For each variable it returns: Quantity and percentage of zeros (q_zeros and p_zeros respectevly). Same metrics for NA values (q_NA/p_na), and infinite values (q_inf/p_inf). Last two columns indicates data type and quantity of unique values. This function print and return the results.

```
#install.packages("funModeling")
library(funModeling)
```

```
## 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
```

```
## funModeling v.1.9.3 :)
## Examples and tutorials at livebook.datascienceheroes.com
##  / Now in Spanish: librovivodecienciadedatos.ai
```

```
df_status(data1)
```

```
##                   variable q_zeros p_zeros q_na p_na q_inf p_inf
type
## 1             Store_number       0    0.00    0    0     0     0
character
## 2           Commodity_Name       0    0.00    0    0     0     0
character
## 3         Base_GTIN_Number       0    0.00    0    0     0     0
character
## 4    Base_GTIN_Description       0    0.00    0    0     0     0
character
## 5 Adjustment_Unit_Quantity  303461   33.98    0    0     0     0
numeric
##   unique
## 1     30
## 2    542
## 3  62568
## 4  53709
## 5   1017
```

Variance and Standard deviation of Adjustment_Unit_Quantity for all stores Standard deviation is sensitive to outliers. A single outlier can raise the standard deviation and in turn, distort the picture of spread.SD is the best measure of spread of an approximately normal distribution. This is not the case when there are extreme values in a distribution or when the distribution is skewed, in these situations interquartile range or semi-interquartile are preferred measures of spread. Interquartile range is the difference between the 25th and 75th centiles.

```
 quantile(data1$Adjustment_Unit_Quantity)
```

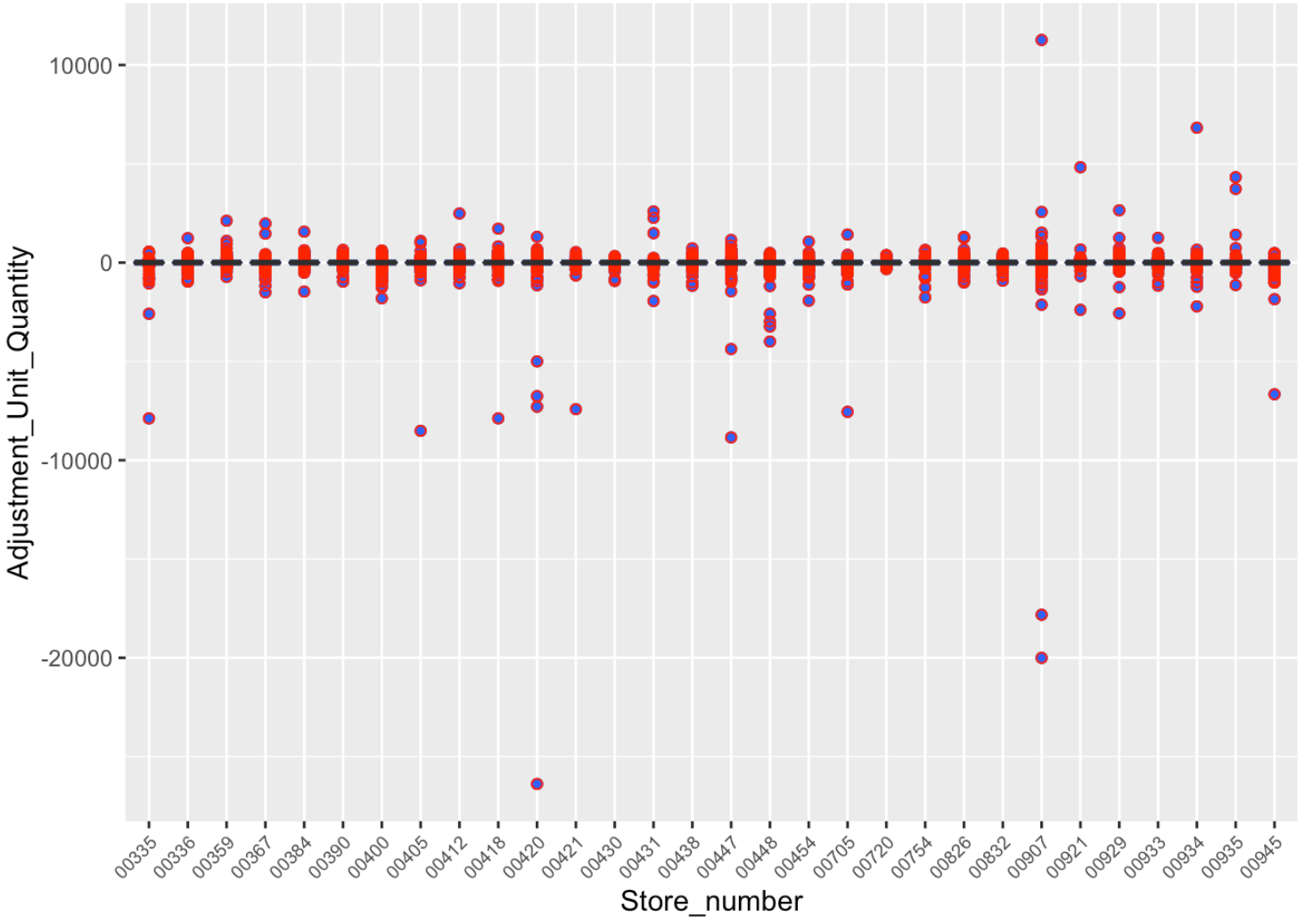
```
##      0%     25%     50%     75%    100%
## -26384      -1       0       1   11268
```

```
IQR(data1$Adjustment_Unit_Quantity)
```

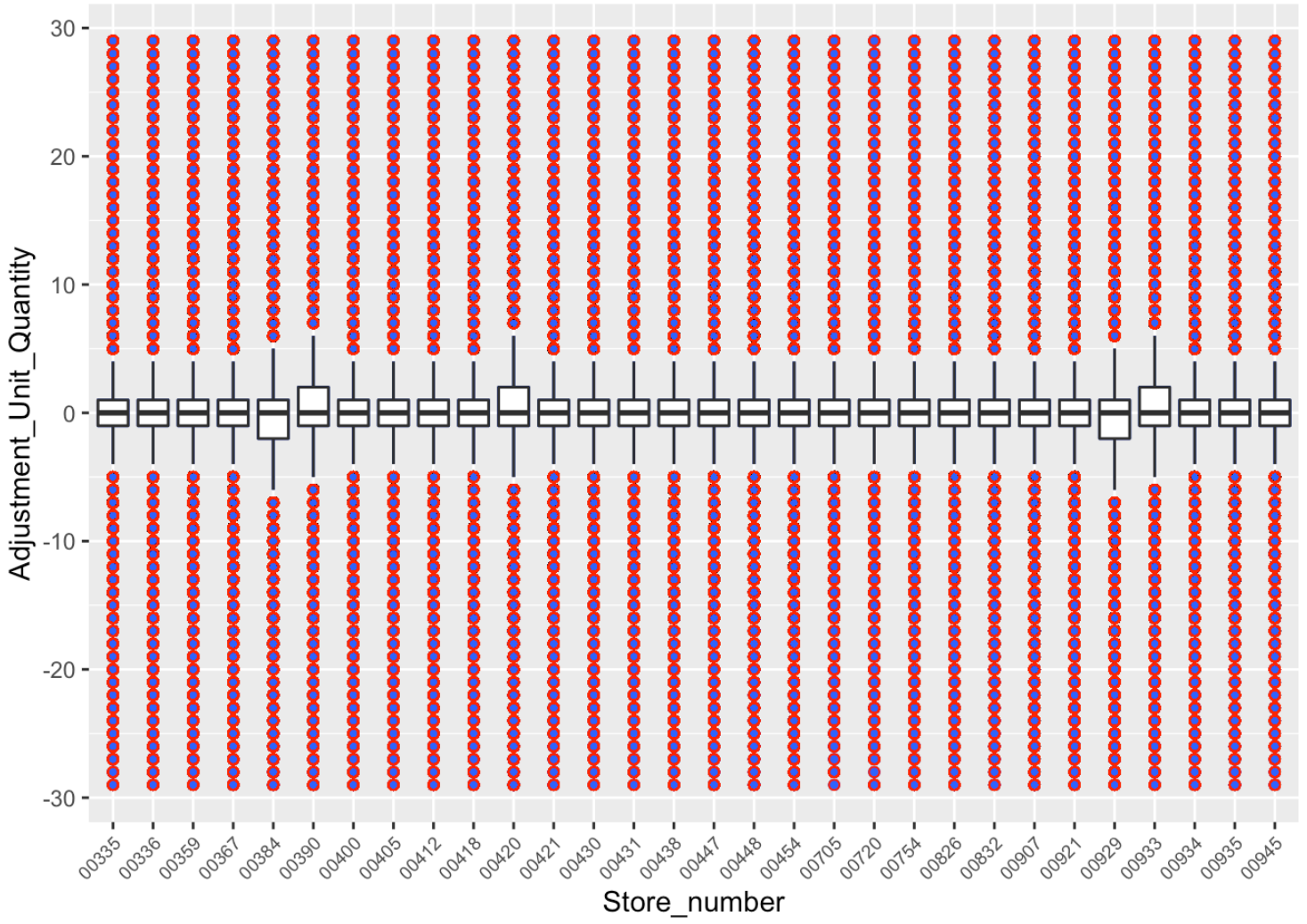
```
## [1] 2
```

Create boxplot to determine outliers

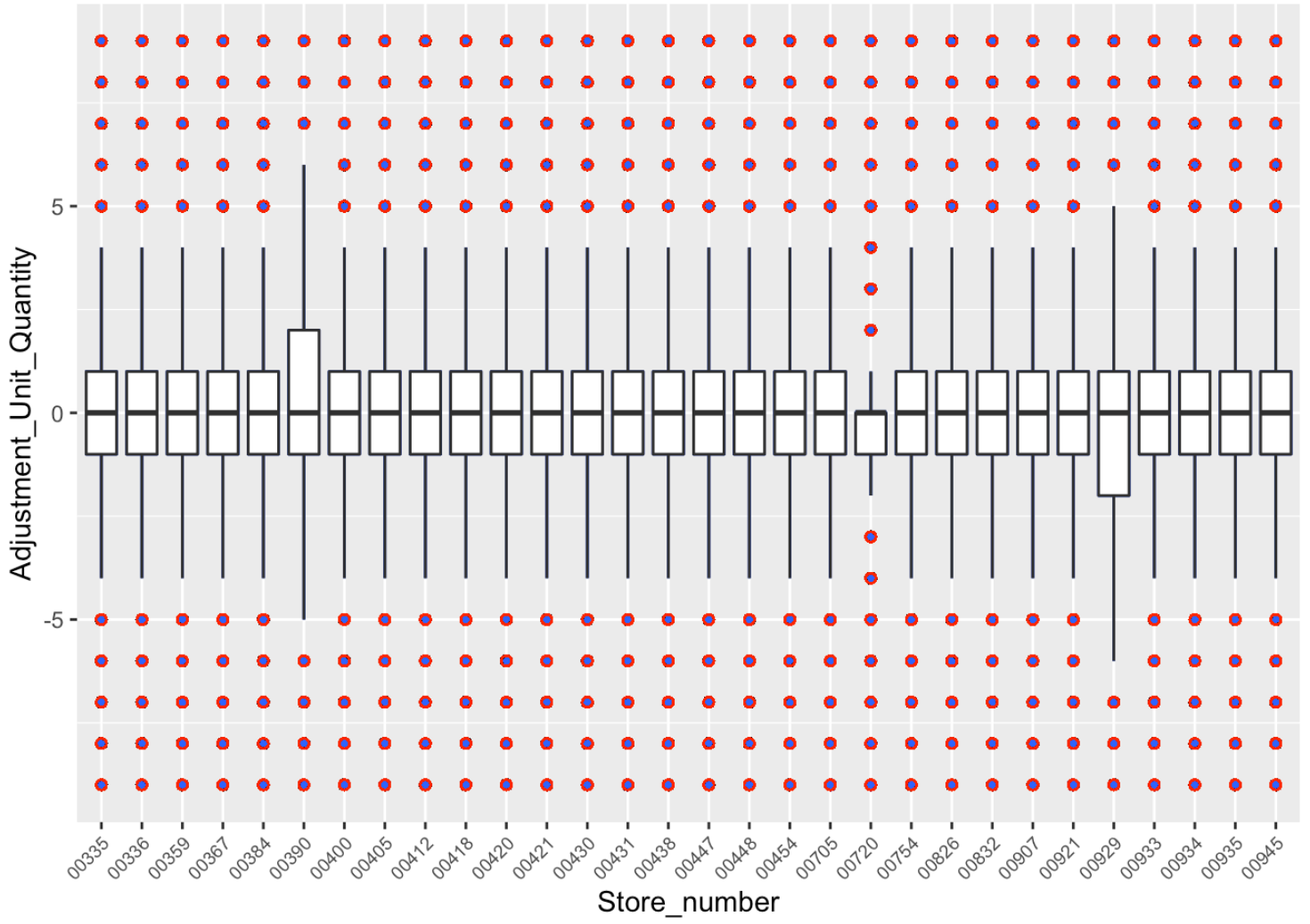
```
 p <- ggplot(data1, aes(Store_number, Adjustment_Unit_Quantity))
p1 <- p + geom_boxplot()
p2 <- p1 + geom_boxplot(fill = "white", colour = "#3366FF")
  p2 + geom_boxplot(outlier.colour = "red", outlier.shape = 1)+ them
e(plot.title = element_text(size =10),axis.text.x = element_text(siz
e =7,angle = 45, hjust = 1))
```



```
#keep observations between 30 to -30
data11 <- data1 %>% filter(Adjustment_Unit_Quantity < 30 & Adjustmen
t_Unit_Quantity > -30)
r <- ggplot(data11, aes(Store_number, Adjustment_Unit_Quantity))
r1 <- r + geom_boxplot()
r2 <- r1 + geom_boxplot(fill = "white", colour = "#3366FF")
r2 + geom_boxplot(outlier.colour = "red", outlier.shape = 1)+ theme(
plot.title = element_text(size =10),axis.text.x = element_text(size
=7,angle = 45, hjust = 1))
```



```
 #keep observations between 10 to -10
data11 <- data1 %>% filter(Adjustment_Unit_Quantity < 10 & Adjustmen
t_Unit_Quantity > -10)
Q <- ggplot(data11, aes(Store_number, Adjustment_Unit_Quantity))
Q1 <- Q + geom_boxplot()
Q2 <- Q1 + geom_boxplot(fill = "white", colour = "#3366FF")
Q2 + geom_boxplot(outlier.colour = "red", outlier.shape = 1)+theme(p
lot.title = element_text(size =10),axis.text.x = element_text(size =
7,angle = 45, hjust = 1))
```



Not all outliers are bad and some should not be deleted. In fact, outliers can be very informative about the subject-area and data collection process. It's important to understand how outliers occur and whether they might happen again as a normal part of the process or study area.

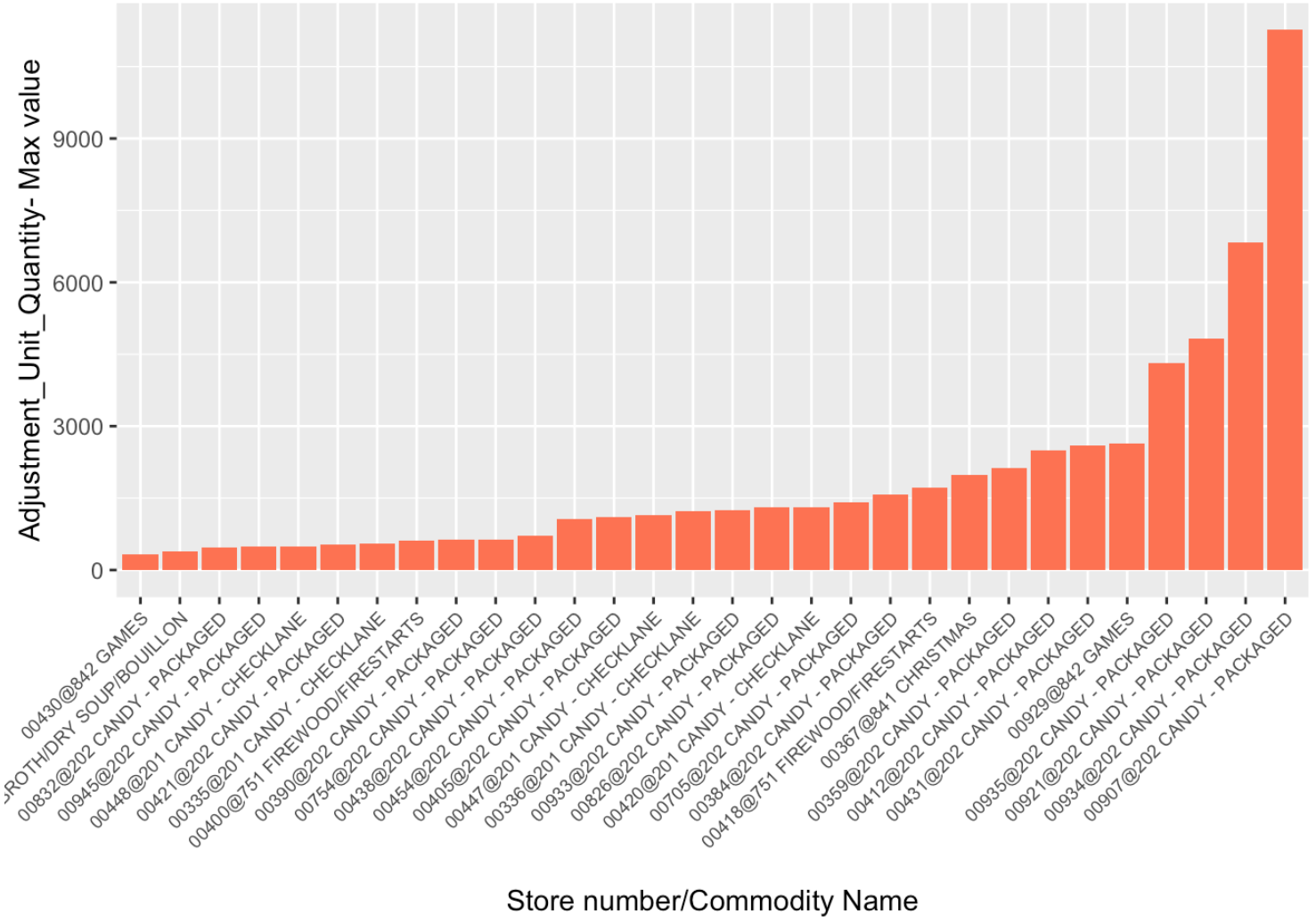
Data Analysis

```
# get maximum, minimum and median for commodity
a <- sqldf("select max(Adjustment_Unit_Quantity)  as maxAdjQty, stor
e_number, commodity_name, Base_GTIN_Number from data1 group by Store
_number")

a1 <- a %>% unite("St_Co", Store_number, Commodity_Name, sep="@")


b <- sqldf("select median(Adjustment_Unit_Quantity)  as maxAdjQty, s
tore_number, commodity_name, Base_GTIN_Number from data1 group by St
ore_number")

b1 <- b %>% unite("St_Co", Store_number, Commodity_Name, sep="@")


c <- sqldf("select min(Adjustment_Unit_Quantity)  as minAdjQty, stor
e_number, commodity_name, Base_GTIN_Number from data1 group by Store
_number")

c1 <- c %>% unite("St_Co", Store_number, Commodity_Name, sep="@")

y <- sqldf("select max(Adjustment_Unit_Quantity)  as maxAdjQty,   mi
n(Adjustment_Unit_Quantity) as minAdjQty, median(Adjustment_Unit_Qua
ntity), store_number from data1 group by Store_number")

y1 <- y %>% gather(Store, Measure, -4)

sqldf("select median(Adjustment_Unit_Quantity) from data1 where Stor
e_number like '%420%'")
```

```
##   median(Adjustment_Unit_Quantity)
## 1                                0
```

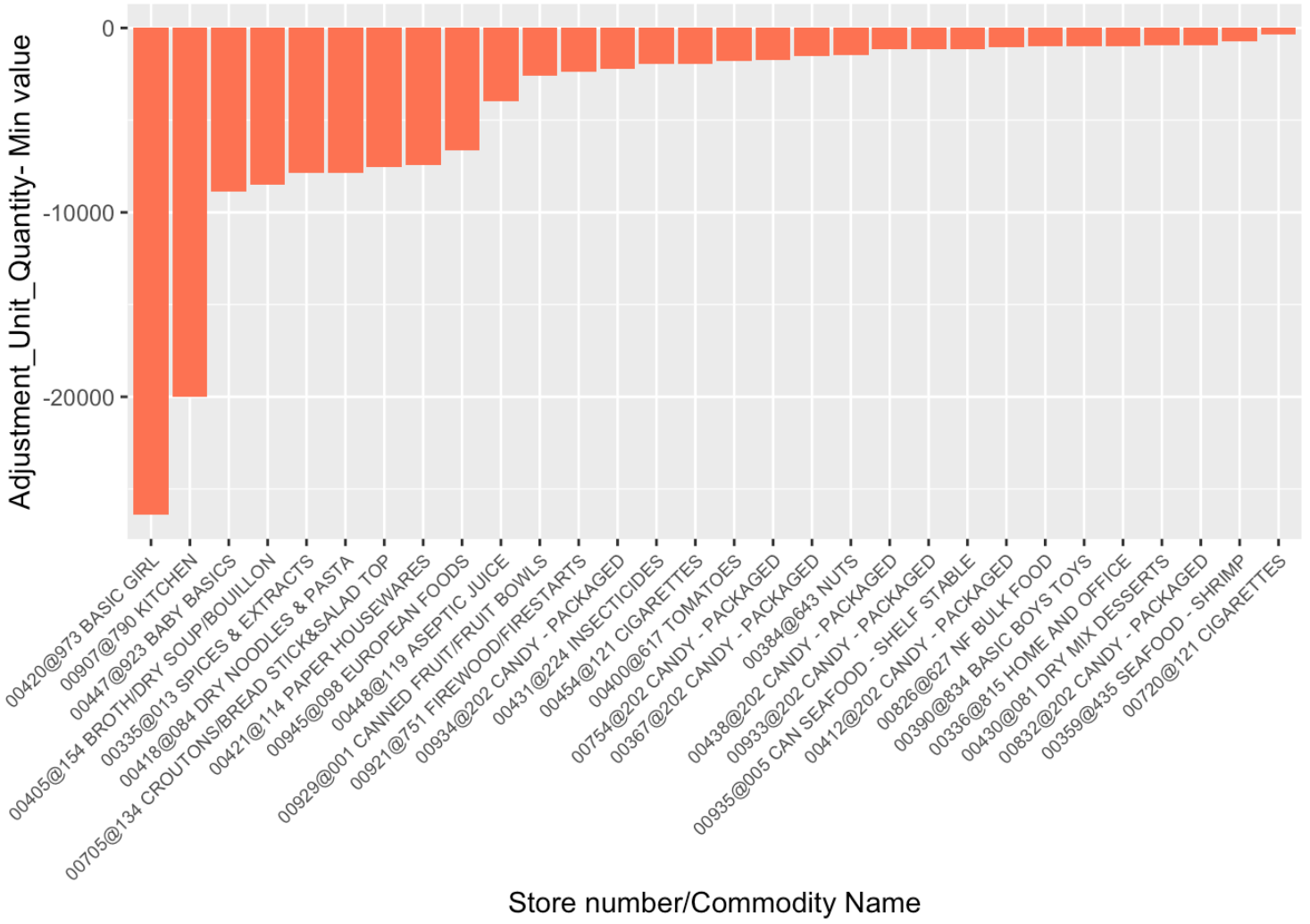
Data VISUALIZATION

```
#boxplot
a1 %>% ggplot(aes(x= reorder(St_Co, maxAdjQty ), y = maxAdjQty)) + g
eom_bar(stat = 'identity', fill = "coral1")  + theme(plot.title = el
ement_text(size =10),axis.text.x = element_text(size =7,angle = 45,
hjust = 1))+xlab("Store number/Commodity Name") + ylab("Adjustment_U
nit_Quantity- Max value")
```



The maximum positive Adjusted Unit Quantity is for 202 Candy-Packaged in Store#907 followed by Store#934, 921 and 935.

```
c1 %>% ggplot(aes(x= reorder(St_Co, minAdjQty ), y = minAdjQty)) + g
eom_bar(stat = 'identity', fill = "coral1")  + theme(plot.title = el
ement_text(size =10),axis.text.x = element_text(size =7,angle = 45,
hjust = 1))+xlab("Store number/Commodity Name") + ylab("Adjustment_U
nit_Quantity- Min value")
```



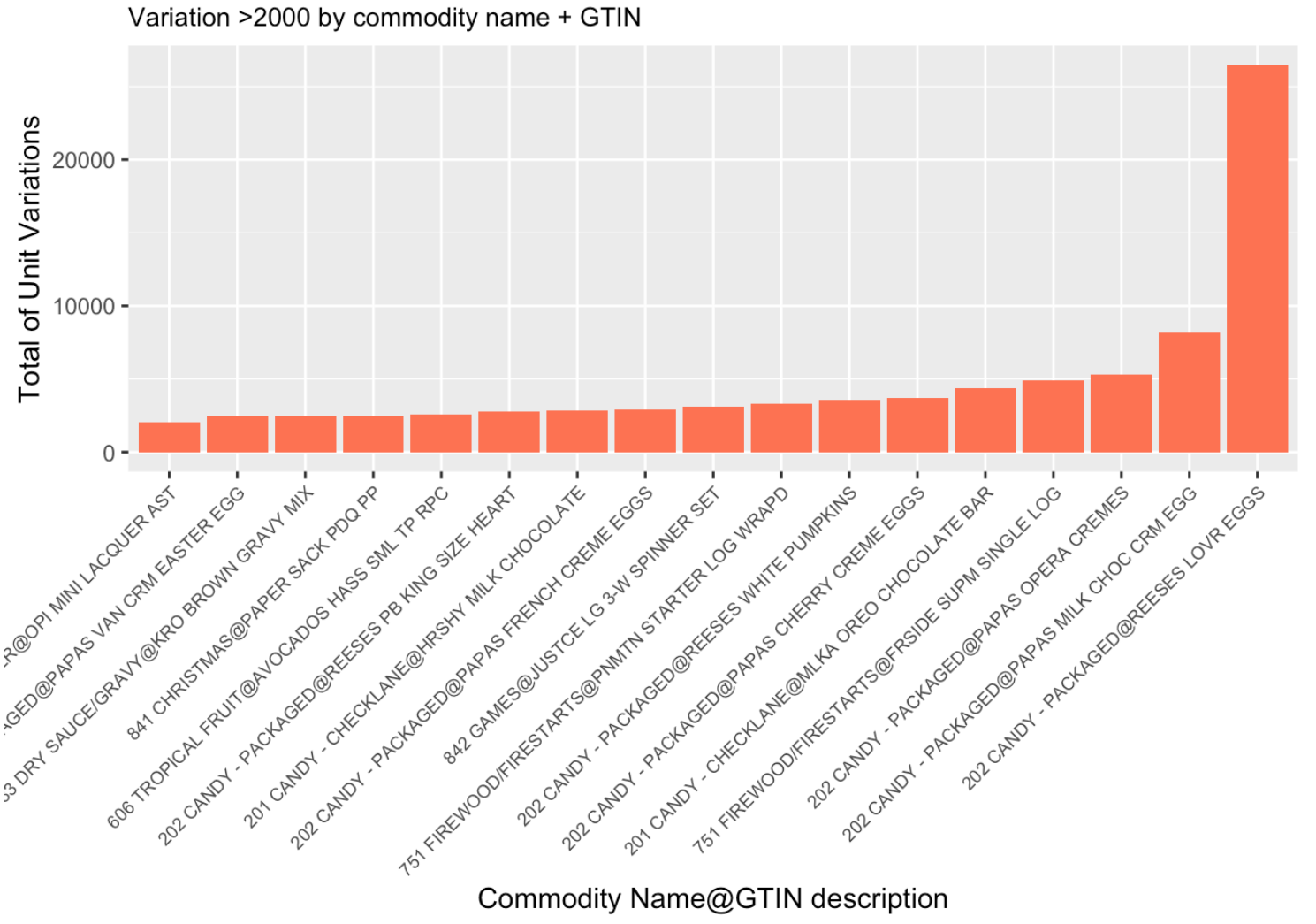
The maximum negative Adjustment Unit Quantity is for
973 Basic Girl in Store#420 followed by
790-Kitchen in Store#907.

Data Analysis

```
#get total variations by commodity + GTIN for all stores
data10 <- data1 %>% group_by(Commodity_Name, Base_GTIN_Description )
%>% summarise(Adj_Qty = sum(Adjustment_Unit_Quantity))

##get total variations by commodity for all stores
data20 <- data1 %>% group_by(Commodity_Name  ) %>% summarise(Adj_Qty
= sum(Adjustment_Unit_Quantity))

#get total variations by Store+ commodity
data30 <- data1 %>% group_by(Store_number,Commodity_Name) %>% summar
ise(Adj_Qty = sum(Adjustment_Unit_Quantity))
```

Data Visualization-total variations by commodity + GTIN for all stores

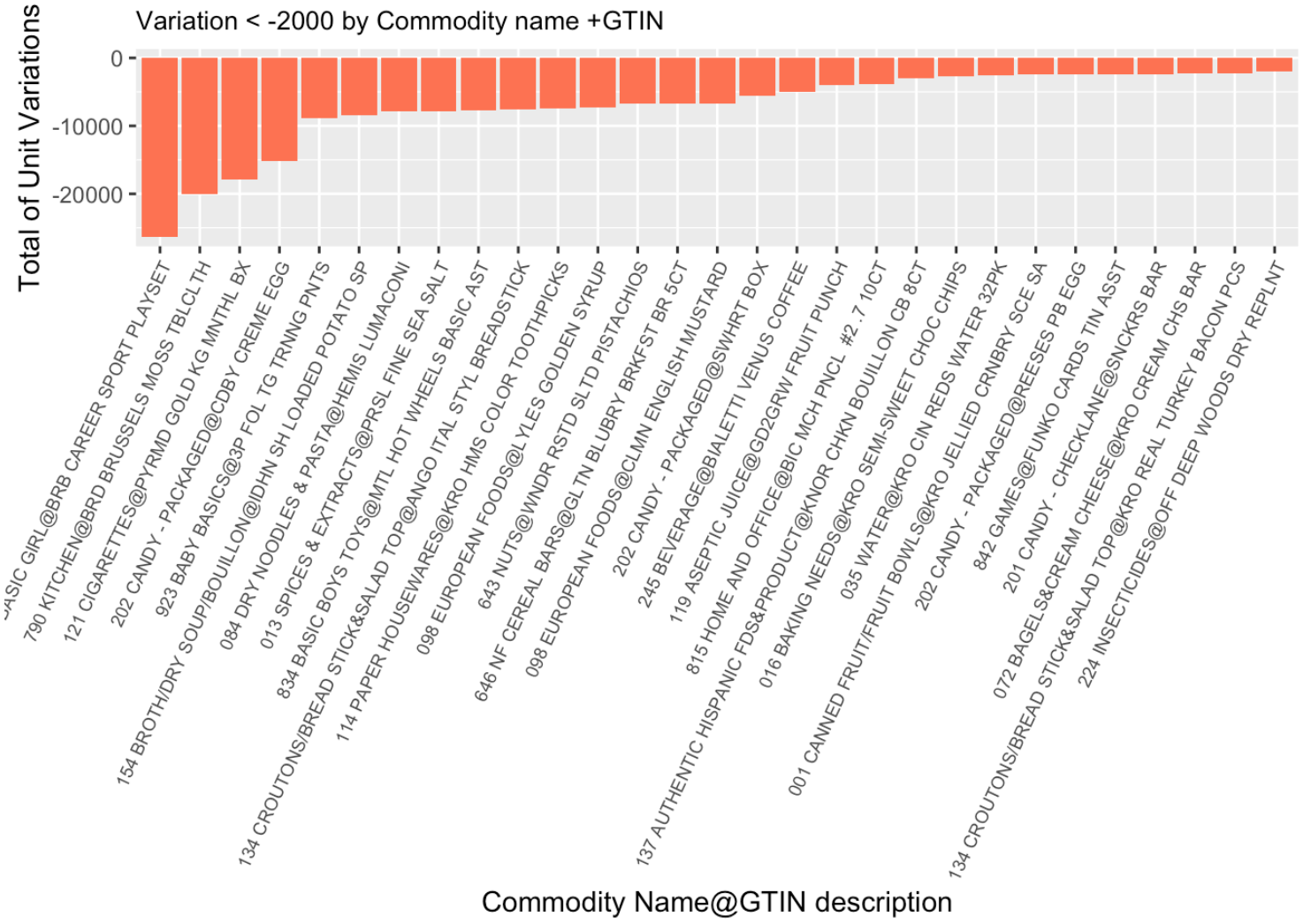
```
data10_01 <- data10 %>% unite("Co_GT", Commodity_Name, Base_GTIN_Des
cription, sep="@")
data11 <- data10_01 %>% filter( Adj_Qty > 2000 )

data11 %>%
ggplot(aes(x=reorder(Co_GT, Adj_Qty, FUN = sum), y=Adj_Qty)) + geom_
bar(stat = 'identity', fill = "coral1") + ggtitle("Variation >2000 b
y commodity name + GTIN") + theme(plot.title = element_text(size =10
),axis.text.x = element_text(size =7,angle = 45, hjust = 1))  + xlab
("Commodity Name@GTIN description") +ylab("Total of Unit Variations"
)
```



The maximum positive total of all unit variations when grouped by commodity name and GTIN is for
202 CANDY - PACKAGED@REESES (mailto:PACKAGED@REESES) LOVR EGGS

Data Visualization-total variations by commodity + GTIN for all stores

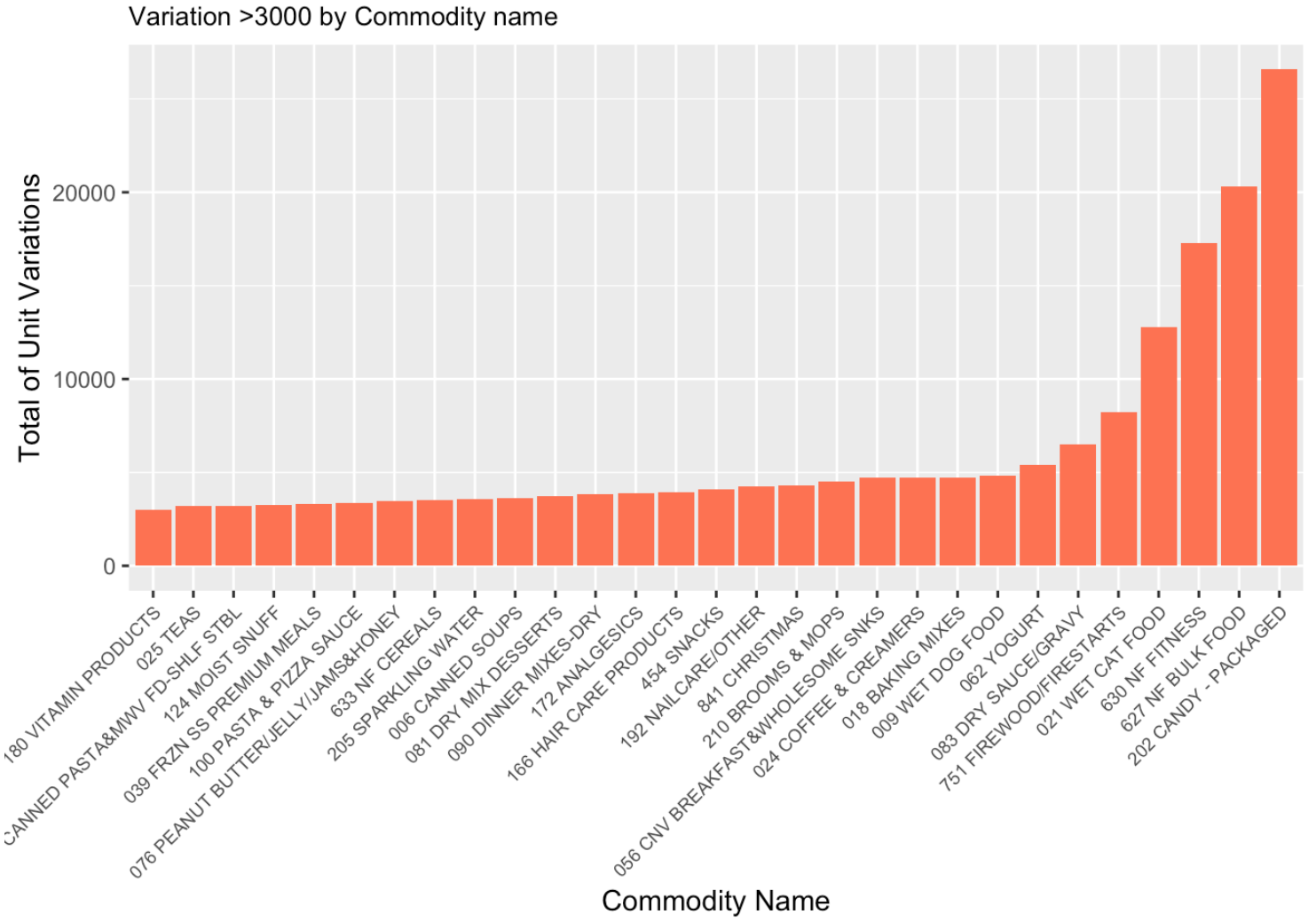
```
data12 <- data10_01 %>% filter( Adj_Qty < -2000)

 data12 %>%
ggplot(aes(x = reorder(Co_GT, Adj_Qty, FUN = sum), y= Adj_Qty)) + ge
om_bar(stat = 'identity', fill = "coral1") + ggtitle("Variation < -2
000 by Commodity name +GTIN")  + theme(plot.title = element_text(siz
e =10),axis.text.x = element_text(size =7,angle = 65, hjust = 1))  +
xlab("Commodity Name@GTIN description")  +ylab("Total of Unit Variat
ions")
```



The max negative value of total unit variations when grouped by comodity name and GTIN is for
973 BASIC GIRL@BRB (mailto:GIRL@BRB) CAREER SPORT PLAYSET, followed by
790 KITCHEN@BRD (mailto:KITCHEN@BRD) BRUSSELS MOSS TBLCLTH,
121 CIGARETTES@PYRMD (mailto:CIGARETTES@PYRMD) GOLD KG MNTHL BX and
202 CANDY - PACKAGED@CDBY (mailto:PACKAGED@CDBY) CREME EGG

Data Visualization-total variations by commodity for all stores

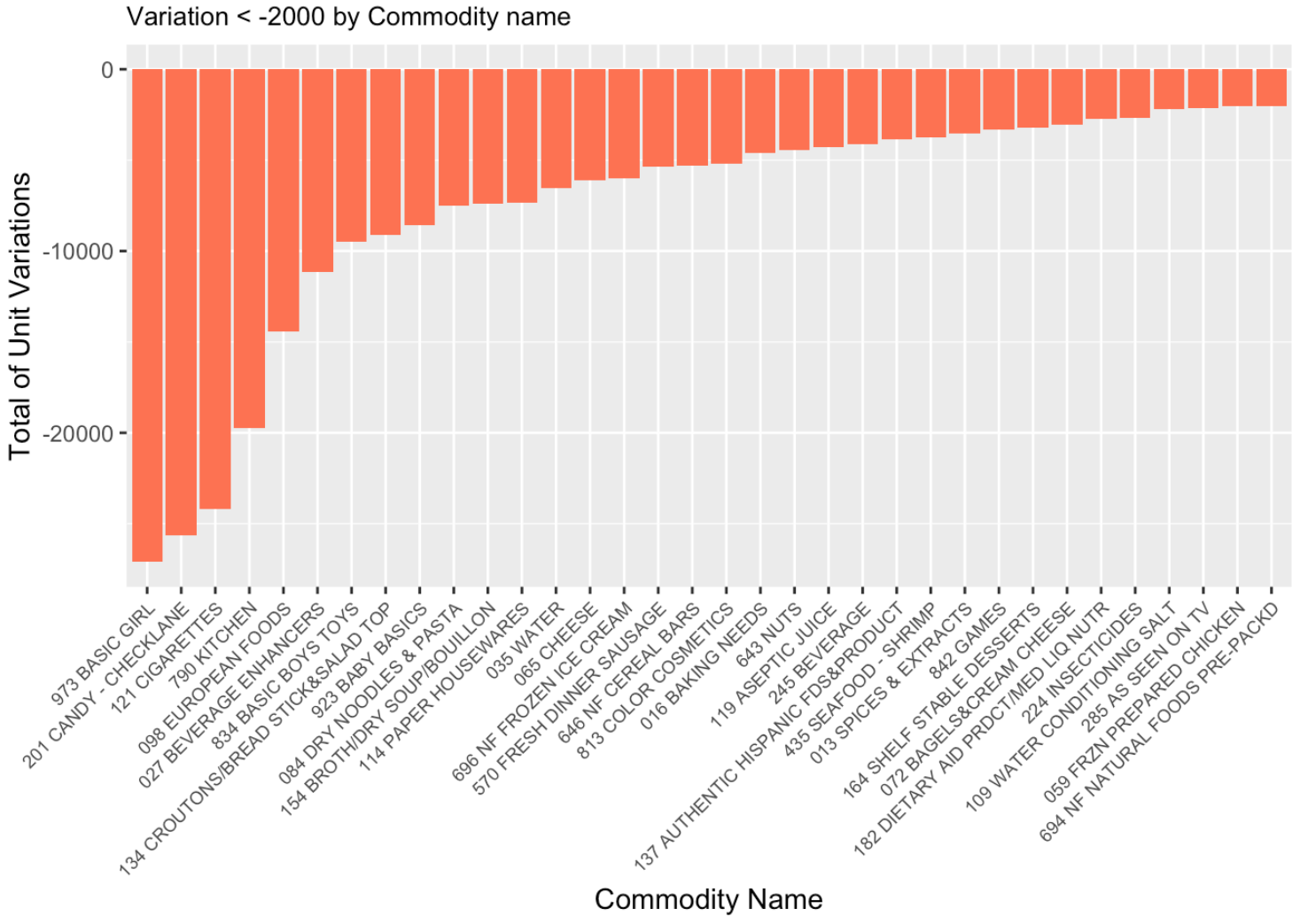
```
data21 <- data20 %>% filter( Adj_Qty >3000)

data21 %>%
ggplot(aes(x=reorder(Commodity_Name, Adj_Qty, FUN = sum), y=Adj_Qty)
) + geom_bar(stat = 'identity', fill = "coral1") + ggtitle("Variatio
n >3000 by Commodity name") + theme(plot.title = element_text(size =
10),axis.text.x = element_text(size =7,angle = 45, hjust = 1))  + xl
ab("Commodity Name") +ylab("Total of Unit Variations")
```



The maximum positive of total variations for commodity is for
202 CANDY - PACKAGED followed by
627 NF BULK FOOD,
630 NF FITNESS and
021 WET CAT FOOD.

Data Visualization-total variations by commodity for all stores

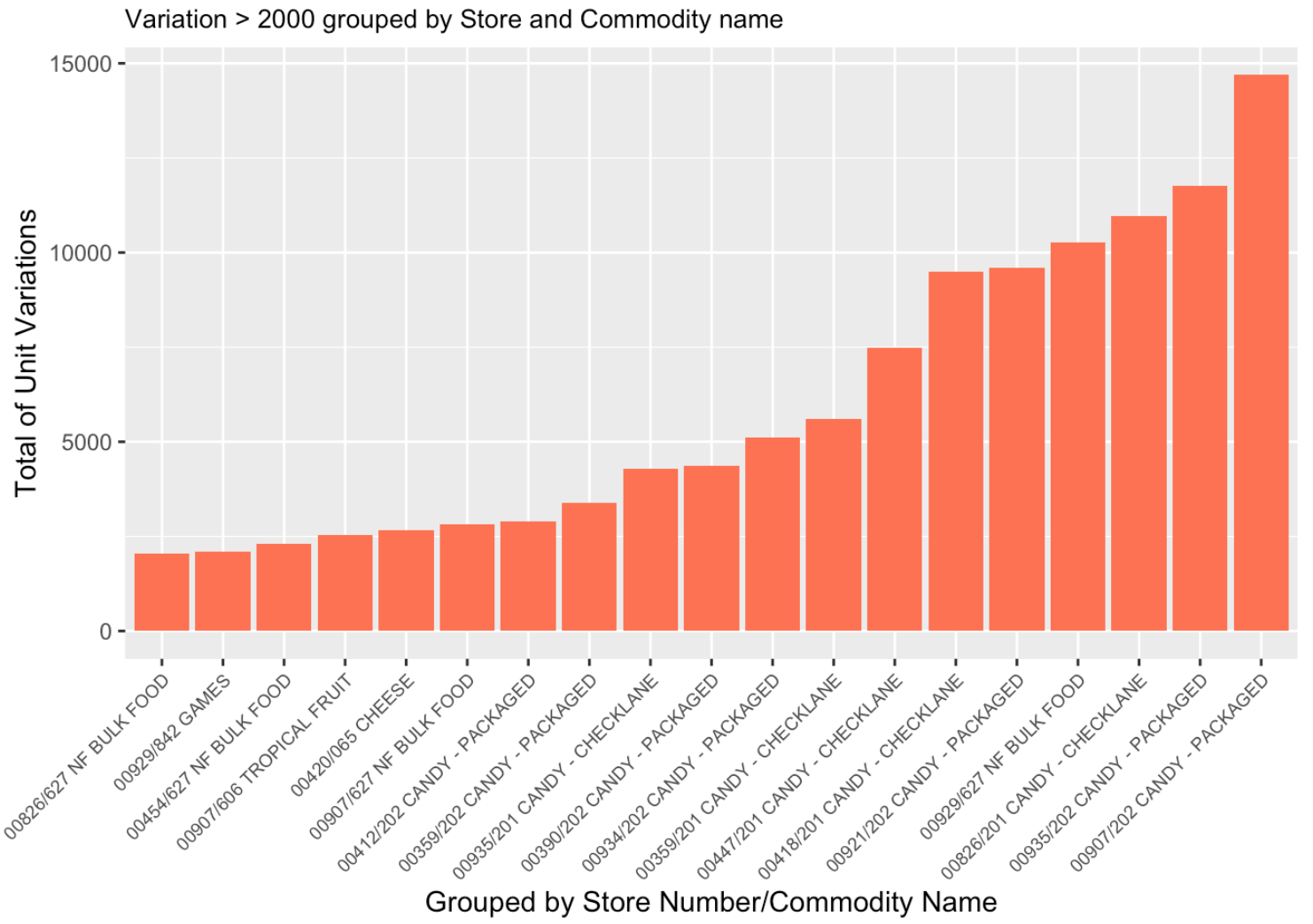
```
data22 <- data20 %>% filter( Adj_Qty < -2000)

 data22 %>%
ggplot(aes(x = reorder(Commodity_Name, Adj_Qty, FUN = sum), y= Adj_Q
ty)) + geom_bar(stat = 'identity', fill = "coral1") + ggtitle("Varia
tion < -2000 by Commodity name")  + theme(plot.title = element_text(
size =10),axis.text.x = element_text(size =7,angle = 45, hjust = 1))
+ xlab("Commodity Name") +ylab("Total of Unit Variations")
```



The maximum negative of total variations for commodity is for
973 BASIC GIRL followed by
201 CANDY - CHECKLANE,
121 CIGARETTES,
790 KITCHEN,
098 EUROPEAN FOODS and
027 BEVERAGE ENHANCERS.

Data Visualization-total variations by Store +commodity for all stores

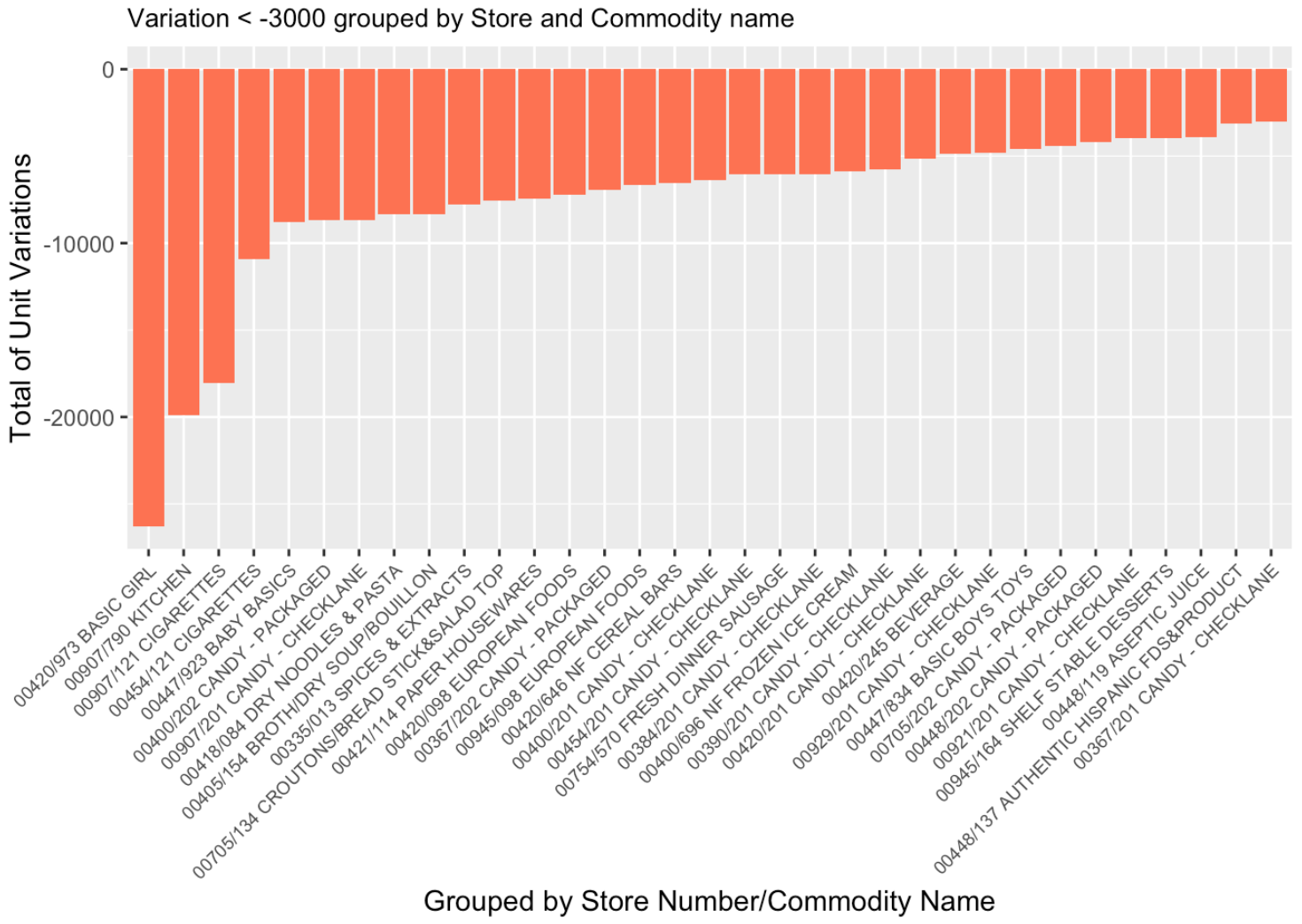
```
data30_01 <- data30 %>% unite("St_Co", Store_number, Commodity_Name,
sep="/")
data31 <- data30_01 %>% filter( Adj_Qty > 2000 )

data31%>%
ggplot(aes(x=reorder(St_Co, Adj_Qty, FUN = sum), y=Adj_Qty)) + geom_
bar(stat = 'identity', fill = "coral1") + ggtitle("Variation > 2000
grouped by Store and Commodity name")   + xlab("Grouped by Store Num
ber/Commodity Name") +ylab("Total of Unit Variations")  + theme(plot
.title = element_text(size =10),axis.text.x = element_text(size =7,a
ngle = 45, hjust = 1))
```



The max positive value of total unit variations when grouped by Store number and comodity name is for 00907/202 CANDY - PACKAGED,
00935/202 CANDY - PACKAGED,
00826/201 CANDY - CHECKLANE and
00929/627 NF BULK FOOD

Data Visualization-total variations by Store + commodity for all stores

```
data30_01 <- data30 %>% unite("St_Co", Store_number, Commodity_Name,
sep="/")
data32 <- data30_01 %>% filter( Adj_Qty < -3000 )

data32%>%
ggplot(aes(x=reorder(St_Co, Adj_Qty, FUN = sum), y=Adj_Qty)) + geom_
bar(stat = 'identity', fill = "coral1") + ggtitle("Variation < -3000
grouped by Store and Commodity name")   + xlab("Grouped by Store Num
ber/Commodity Name") +ylab("Total of Unit Variations")  + theme(plot
.title = element_text(size =10),axis.text.x = element_text(size =7,a
ngle = 45, hjust = 1))
```



The max negative value of total unit variations when grouped by Store number and comodity name is for
00420/973 BASIC GIRL,
00907/790 KITCHEN,
00907/121 CIGARETTES and
00454/121 CIGARETTES.

Analysis did show that there were large number of outliers with the 25th percentile value of Adjustment Unit quantity as -1 and 75th percentile as 1.

In the retail business, it is good to keep and analyse outliers to assess the demand ofcommodities and to take advantage of high demands and find out why the demands are low. The demands in-turn reflects the effect on the profit and loss for the stores.

Besides individual max and min Unit variations, the sum of all the variations by commodity, or by grouping with store or GTIN, the supplier get a perspective of the variations for a set period and can investigate for the outliers for the cause and the impact on the profit or loss and reducing unnecessary inventory.