

# Eric\_Hirsch\_621\_Assignment\_4

## Predicting Insurance Claims

Eric Hirsch

4/7/2022

## Contents

1. Data Exploration . . . . .	1
A. Summary Statistics . . . . .	1
B. Multicollinearity . . . . .	2
C. Distributions . . . . .	4
B. Distributions . . . . .	16

```
knitr::opts_chunk$set(echo = TRUE, warning = TRUE, message = TRUE)
```

```
df2 <- read.csv("D:\\RStudio\\CUNY_622\\1\\Salesdata_5000.csv")
#df2 <- read.csv("D:\\RStudio\\CUNY_622\\1\\Salesdata_50000.csv")
```

### 1. Data Exploration

**A. Summary Statistics** For this exercise we will examine the 5000 record and 50,000 record datasets from the assignment website.

The datasets contain fabricated sales orders generated by VBA for the purpose of practicing analysis. There are 14 columns, including 7 numeric columns, 5 character and two date. One of the variables is an ID so we drop it. Here is a summary of the remaining 13 variables:

```
df2 <- df2 %>%
  dplyr::select(-Order.ID)

summary(df2)
```

```
##      Region      Country      Item.Type      Sales.Channel
## Length:5000    Length:5000    Length:5000    Length:5000
## Class :character Class :character Class :character Class :character
## Mode  :character Mode  :character Mode  :character Mode  :character
##
##
##      Order.Priority      Order.Date      Ship.Date      Units.Sold
## Length:5000            Length:5000    Length:5000    Min.      : 2
```

```
## Class :character    Class :character    Class :character    1st Qu.:2453
## Mode  :character    Mode  :character    Mode  :character    Median :5123
##                                     Mean  :5031
##                                     3rd Qu.:7576
##                                     Max.   :9999
## Unit.Price          Unit.Cost          Total.Revenue          Total.Cost
## Min.   : 9.33      Min.   : 6.92      Min.   : 65      Min.   : 48
## 1st Qu.: 81.73     1st Qu.: 35.84     1st Qu.: 257417    1st Qu.: 154748
## Median :154.06     Median : 97.44     Median : 779409    Median : 468181
## Mean   :265.75     Mean   :187.49     Mean   :1325738    Mean   : 933093
## 3rd Qu.:437.20     3rd Qu.:263.33     3rd Qu.:1839975    3rd Qu.:1189578
## Max.   :668.27     Max.   :524.96     Max.   :6672676    Max.   :5248025
## Total.Profit
## Min.   : 16.9
## 1st Qu.: 85339.3
## Median : 279095.2
## Mean   : 392644.6
## 3rd Qu.: 565106.4
## Max.   :1726007.5
```

```
str(df2)
```

```
## 'data.frame':    5000 obs. of  13 variables:
## $ Region      : chr  "Central America and the Caribbean" "Central America and the Caribbean" "Eur
## $ Country     : chr  "Antigua and Barbuda " "Panama" "Czech Republic" "North Korea" ...
## $ Item.Type    : chr  "Baby Food" "Snacks" "Beverages" "Cereal" ...
## $ Sales.Channel : chr  "Online" "Offline" "Offline" "Offline" ...
## $ Order.Priority: chr  "M" "C" "C" "L" ...
## $ Order.Date   : chr  "12/20/2013" "7/5/2010" "9/12/2011" "5/13/2010" ...
## $ Ship.Date    : chr  "1/11/2014" "7/26/2010" "9/29/2011" "6/15/2010" ...
## $ Units.Sold   : int  552 2167 4778 9016 7542 48 8258 927 8841 9817 ...
## $ Unit.Price   : num  255.3 152.6 47.5 205.7 152.6 ...
## $ Unit.Cost    : num  159.4 97.4 31.8 117.1 97.4 ...
## $ Total.Revenue : num  140915 330641 226716 1854591 1150758 ...
## $ Total.Cost    : num  88000 211152 151893 1055864 734892 ...
## $ Total.Profit  : num  52915 119488 74823 798727 415866 ...
```

```
df2$Item.Type <- factor(df2$Item.Type)
df2$Region <- factor(df2$Region)
```

**B. Multicollinearity** We suspect a high degree of multicollinearity among the numeric variables, since they are components of each other - for example, total profits is made of costs and revenues, while revenues are determined by prices and volume. We also may assume that order and shipping dates are related, and country and region will also be related.

The heatmap below shows the multicollinearity of the economic variables.

```
df2Num <- df2 %>%
  dplyr::select_if(is.numeric)

z <- EHExplore_Multicollinearity(df2Num, printCorrs = TRUE, title="Multicollinearity Among Economic Variables")

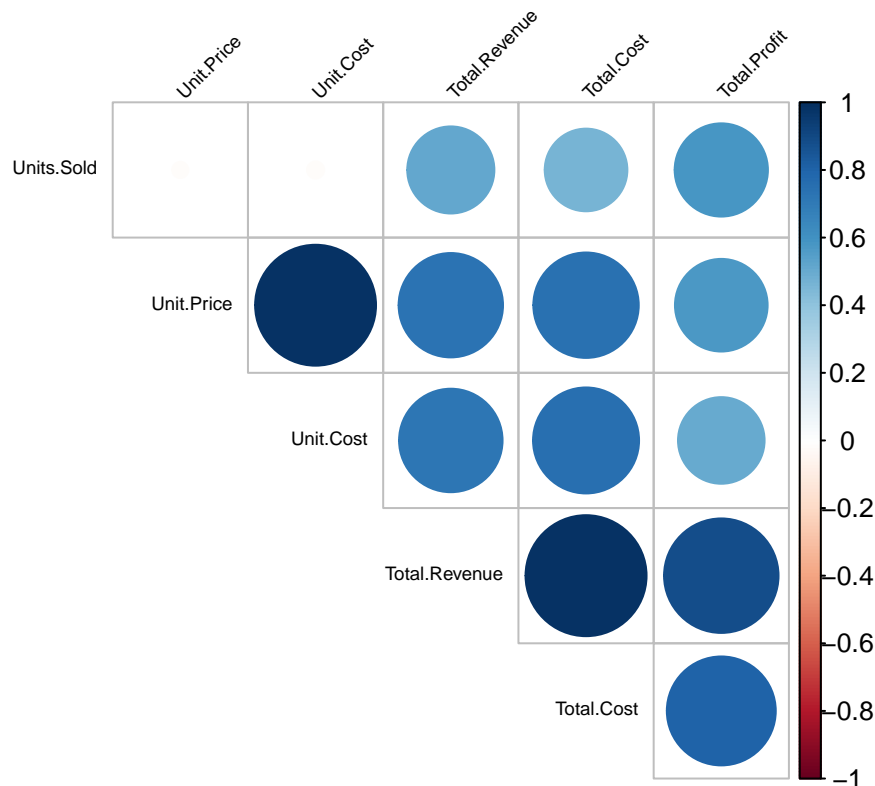
## corplot 0.92 loaded
```

```

##          Units.Sold  Unit.Price  Unit.Cost  Total.Revenue  Total.Cost
## Units.Sold      1.00000000 -0.01749167 -0.01971201      0.5118209  0.4610137
## Unit.Price     -0.01749167  1.00000000  0.98623095      0.7350631  0.7496609
## Unit.Cost      -0.01971201  0.98623095  1.00000000      0.7226761  0.7581004
## Total.Revenue   0.51182089  0.73506309  0.72267611      1.0000000  0.9878272
## Total.Cost      0.46101374  0.74966094  0.75810043      0.9878272  1.0000000
## Total.Profit    0.58641579  0.57902433  0.50593567      0.8839900  0.8005063
##
##          Total.Profit
## Units.Sold      0.5864158
## Unit.Price      0.5790243
## Unit.Cost       0.5059357
## Total.Revenue   0.8839900
## Total.Cost      0.8005063
## Total.Profit    1.0000000

```

## Multicollinearity Among Economic Variables



There are many different strategies we can take with this issue, including ignoring it. We choose, for now, to retain Total Profit (as it summarizes most of the others), and, because the same profit may come from high revenue and high costs or low revenue and low costs, we include Unit Cost as well. Unit cost has the lowest correlation with Total Profit of all the predictors ( $r=0.51$ ).

As for dates, we convert order date to an integer representing the number of days that have passed since 1/1/2000. We also create a new variable, Order.Lag, since the difference between order date and shipping date might be predictive.

Finally, we eliminate country and retain region. This leaves us a dataframe of 8 variables.

```
df2$Order.Date <- as.Date(df2$Order.Date, format="%m/%d/%Y")
df2$Ship.Date <- as.Date(df2$Ship.Date, format="%m/%d/%Y")

df2$Order.Lag <- as.integer(df2$Ship.Date-df2$Order.Date)
df2$OrderDaysSince2000 <- as.integer(df2$Ship.Date-as.Date("2000-01-01"))

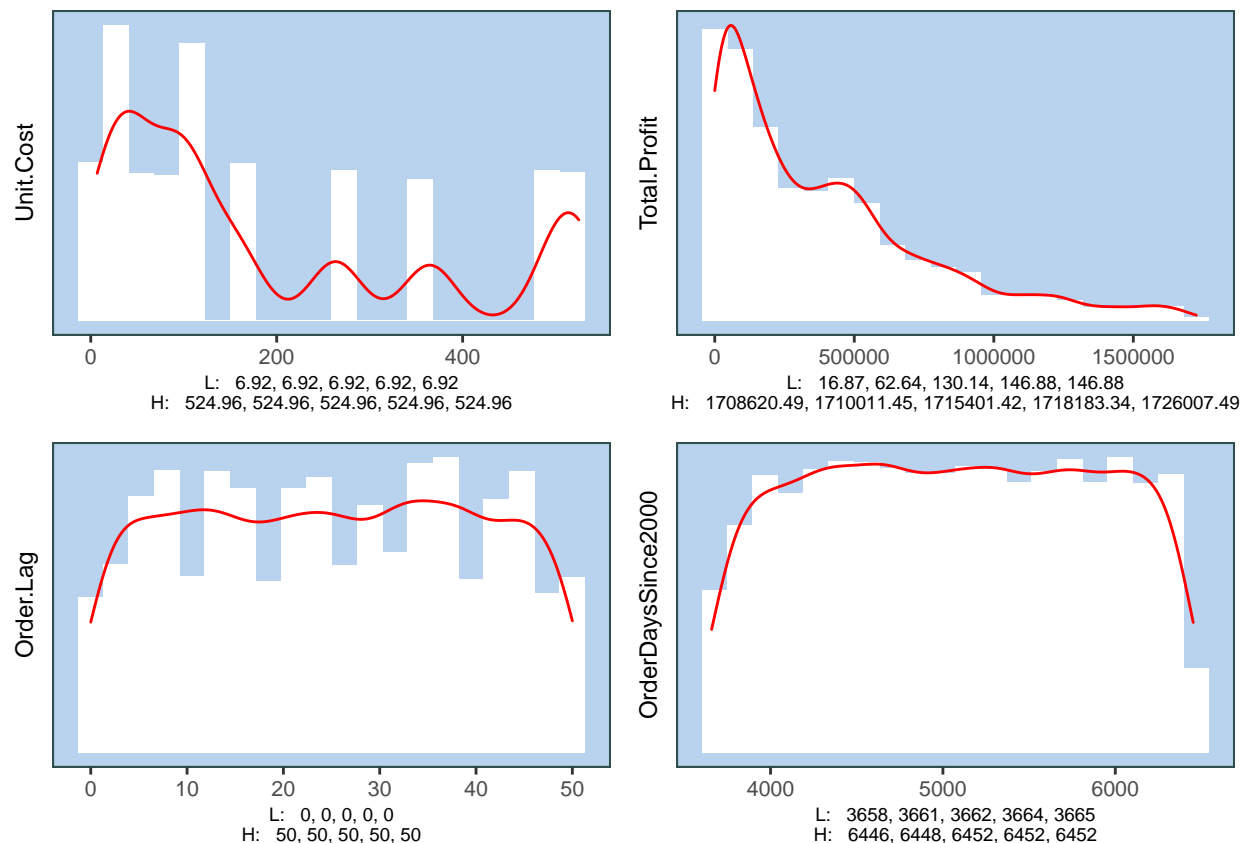
df3 <- df2 %>%
  dplyr::select(-Order.Date, -Ship.Date, -Country, -Unit.Price, -Total.Revenue, -Total.Cost, -Units.S

summary(df3)
```

```
##              Region              Item.Type
## Asia              : 719 Beverages        : 447
## Australia and Oceania : 416 Fruits         : 447
## Central America and the Caribbean: 534 Baby Food      : 445
## Europe              :1330 Cosmetics      : 424
## Middle East and North Africa : 610 Household      : 424
## North America        : 106 Office Supplies: 420
## Sub-Saharan Africa    :1285 (Other)         :2393
## Sales.Channel      Order.Priority      Unit.Cost      Total.Profit
## Length:5000      Length:5000      Min.       : 6.92      Min.       : 16.9
## Class :character  Class :character  1st Qu.: 35.84      1st Qu.: 85339.3
## Mode  :character  Mode  :character  Median : 97.44      Median : 279095.2
##                                     Mean  :187.49      Mean  : 392644.6
##                                     3rd Qu.:263.33      3rd Qu.: 565106.4
##                                     Max.   :524.96      Max.   :1726007.5
##
##      Order.Lag      OrderDaysSince2000
## Min.       : 0.00      Min.       :3658
## 1st Qu.:12.00      1st Qu.:4388
## Median :25.00      Median :5066
## Mean      :25.05      Mean      :5066
## 3rd Qu.:38.00      3rd Qu.:5754
## Max.      :50.00      Max.      :6452
##
```

**C. Distributions** When we examine the distributions of the numeric variables. We find that Total profit is highly skewed, total Cost is somewhat skewed, and the date variables are relatively uniform. There are many gaps in the cost distribution. We may consider doing a log transformation of profit if need be. Since the data is fabricated, the uniformity of the date distributions suggests to me that these dates are just pulled randomly from a uniform distribution and won't be useful.

```
a <- EHSummarize_SingleColumn_Histograms(df3)
grid.arrange(grobs=a[c(1:4)])
```



#### #### D. Relationships

We can run a regression on total profit just to get an idea of some of the relationships between the numeric and categorical variables. We can see from this exploration that item types are strongly correlated with profits, as are medium priority items, but nothing else is. Unit cost could not be calculated because of singularities, which might also be an artifact of the fabrication process. We know Unit Cost is not fully correlated with Total Profit, so it must be fully correlated when in conjunction with other variables.

```
x <- lm(Total.Profit ~., data=df3)
summary(x)
```

```
##
## Call:
## lm(formula = Total.Profit ~ ., data = df3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -869368 -146888   1874  141660  847247
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.970e+05  3.153e+04  15.762  < 2e-16
## RegionAustralia and Oceania    3.597e+03  1.687e+04   0.213  0.8312
## RegionCentral America and the Caribbean -9.154e+02  1.565e+04  -0.058  0.9534
## RegionEurope    -1.667e+04  1.268e+04  -1.314  0.1889
## RegionMiddle East and North Africa  -1.396e+04  1.508e+04  -0.926  0.3545
## RegionNorth America   -3.254e+04  2.850e+04  -1.142  0.2536
```

```

## RegionSub-Saharan Africa          9.604e+03  1.277e+04   0.752   0.4519
## Item.TypeBeverages                -4.117e+05  1.834e+04 -22.452 < 2e-16
## Item.TypeCereal                   -4.083e+04  1.907e+04  -2.141   0.0323
## Item.TypeClothes                  -1.123e+05  1.904e+04  -5.896  3.98e-09
## Item.TypeCosmetics                 3.893e+05  1.857e+04  20.958 < 2e-16
## Item.TypeFruits                   -4.763e+05  1.832e+04 -25.993 < 2e-16
## Item.TypeHousehold                 3.312e+05  1.858e+04  17.826 < 2e-16
## Item.TypeMeat                     -2.166e+05  1.887e+04 -11.475 < 2e-16
## Item.TypeOffice Supplies           1.450e+05  1.863e+04   7.784  8.49e-15
## Item.TypePersonal Care             -3.602e+05  1.868e+04 -19.283 < 2e-16
## Item.TypeSnacks                   -2.236e+05  1.888e+04 -11.846 < 2e-16
## Item.TypeVegetables               -1.665e+05  1.874e+04  -8.888 < 2e-16
## Sales.ChannelOnline               -4.564e+03  7.751e+03  -0.589   0.5560
## Order.PriorityH                    1.026e+04  1.107e+04   0.927   0.3542
## Order.PriorityL                   -3.084e+03  1.119e+04  -0.276   0.7828
## Order.PriorityM                    2.292e+04  1.099e+04   2.085   0.0371
## Unit.Cost                          NA          NA          NA          NA
## Order.Lag                          1.230e+02  2.655e+02   0.463   0.6431
## OrderDaysSince2000                -2.448e+00  4.889e+00  -0.501   0.6165
##
## (Intercept)                        ***
## RegionAustralia and Oceania
## RegionCentral America and the Caribbean
## RegionEurope
## RegionMiddle East and North Africa
## RegionNorth America
## RegionSub-Saharan Africa
## Item.TypeBeverages                ***
## Item.TypeCereal                    *
## Item.TypeClothes                   ***
## Item.TypeCosmetics                 ***
## Item.TypeFruits                    ***
## Item.TypeHousehold                 ***
## Item.TypeMeat                      ***
## Item.TypeOffice Supplies           ***
## Item.TypePersonal Care              ***
## Item.TypeSnacks                     ***
## Item.TypeVegetables                ***
## Sales.ChannelOnline
## Order.PriorityH
## Order.PriorityL
## Order.PriorityM                    *
## Unit.Cost
## Order.Lag
## OrderDaysSince2000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273500 on 4976 degrees of freedom
## Multiple R-squared:  0.4922, Adjusted R-squared:  0.4899
## F-statistic: 209.7 on 23 and 4976 DF, p-value: < 2.2e-16

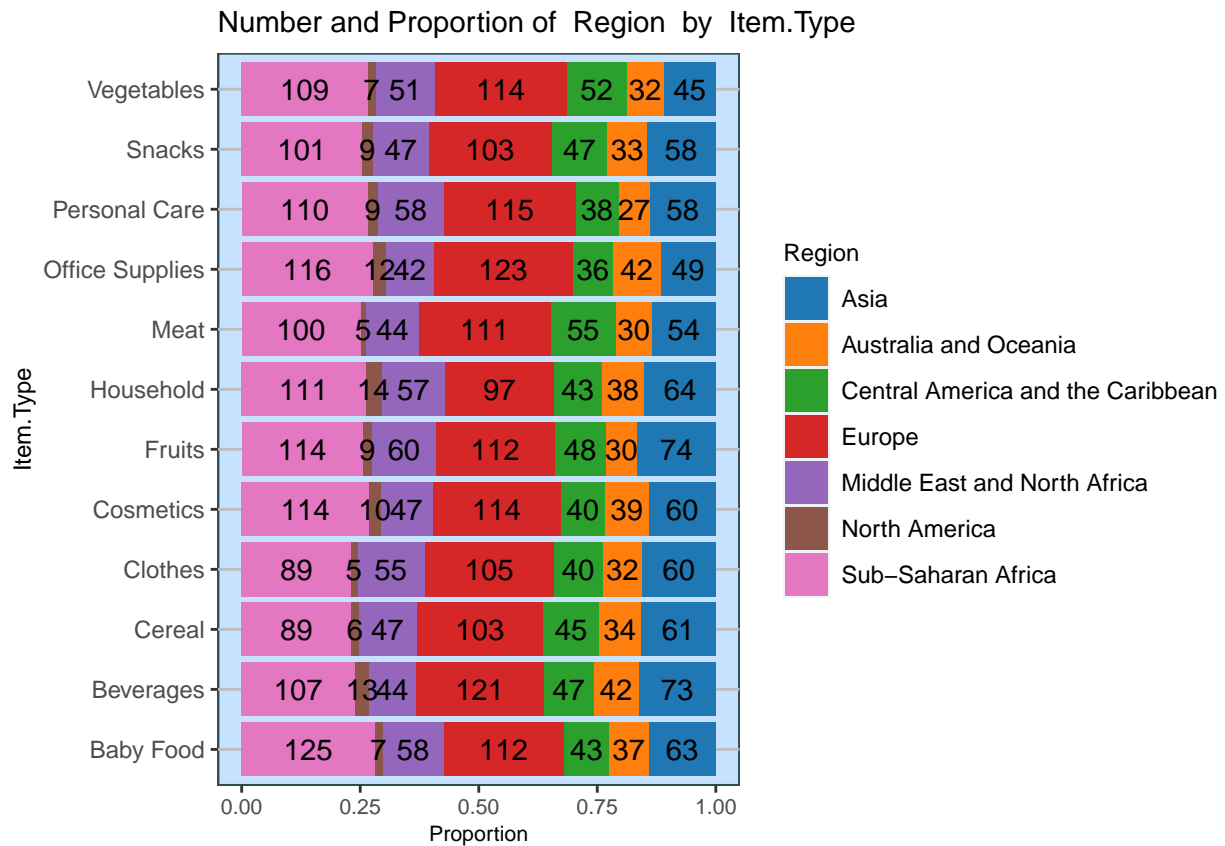
```

This analysis suggests that Item Type may be the most reasonable class to predict. However, we will keep region may be correlated with some of the other variables. We can test this conjecture with some further

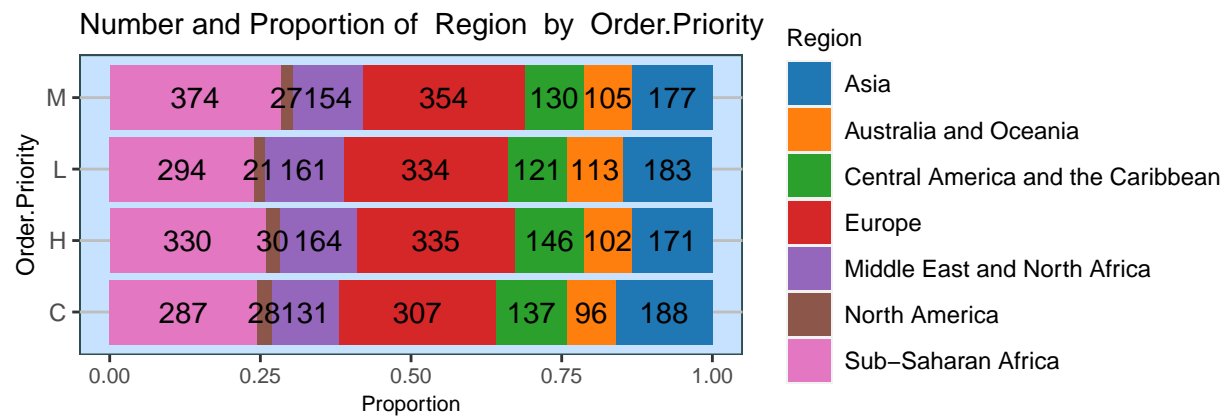
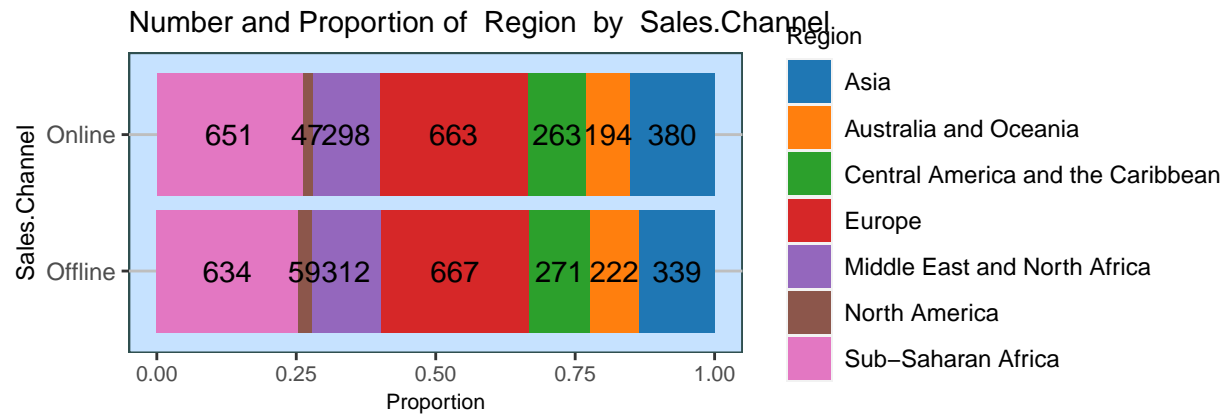
analysis.

Bar charts and boxplots show relatively little relationship between region and item type, sales channel, and order priority.

```
a <- EHExplore_TwoCategoricalColumns_Barcharts(df3, "Region")
grid.arrange(grobs=a[c(2)])
```

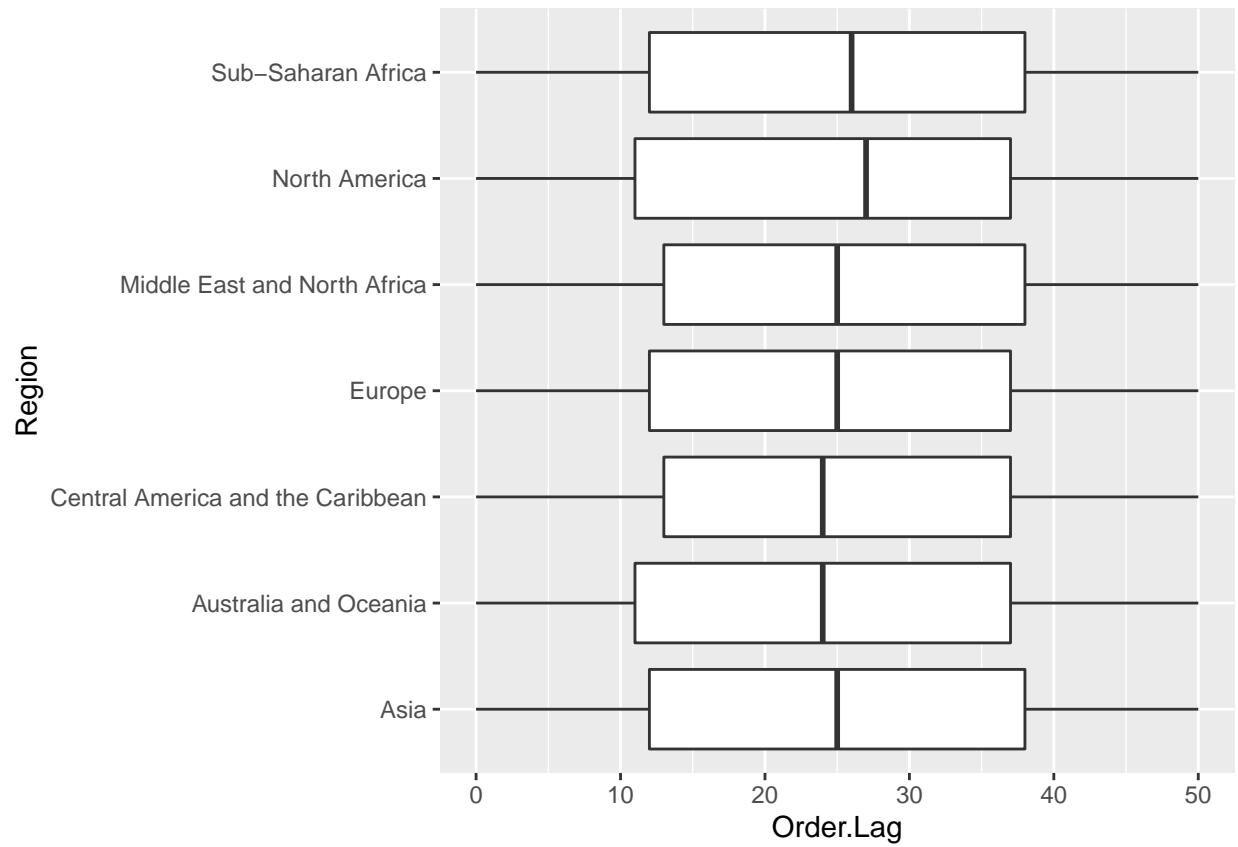


```
grid.arrange(grobs=a[c(3:4)])
```

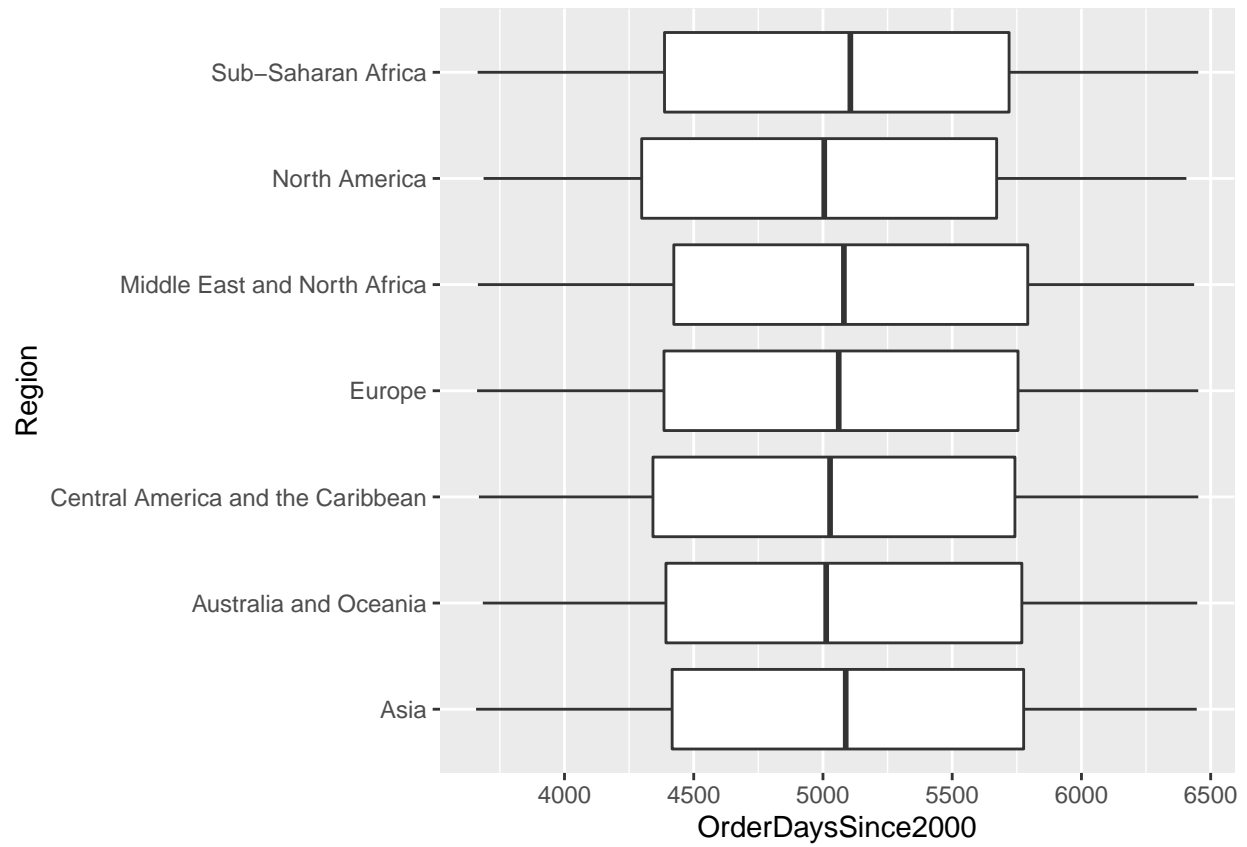


```
ggplot(df2, aes(Region, Order.Lag)) +
  geom_boxplot() +
  coord_flip()
```

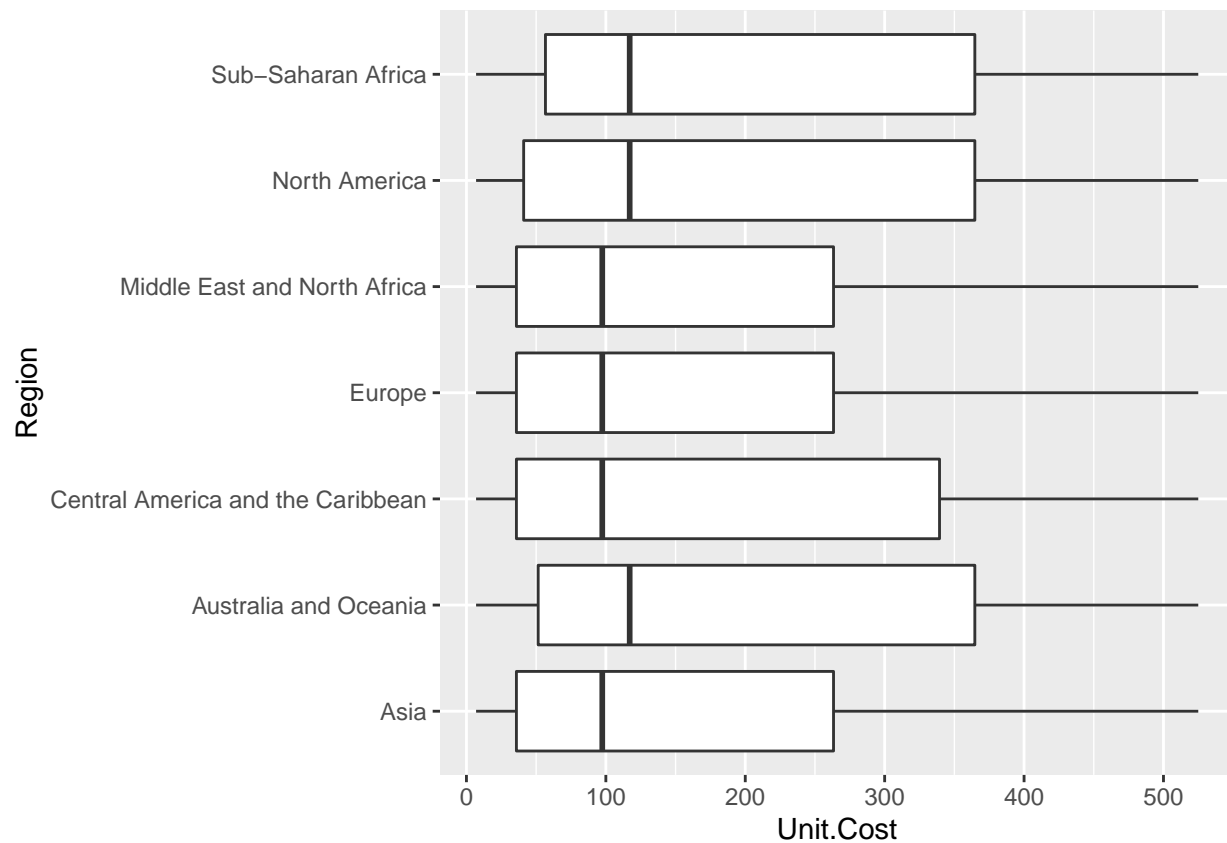




```
ggplot(df2, aes(Region, OrderDaysSince2000)) +  
  geom_boxplot() +  
  coord_flip()
```



```
ggplot(df2, aes(Region, Unit.Cost)) +  
  geom_boxplot() +  
  coord_flip()
```

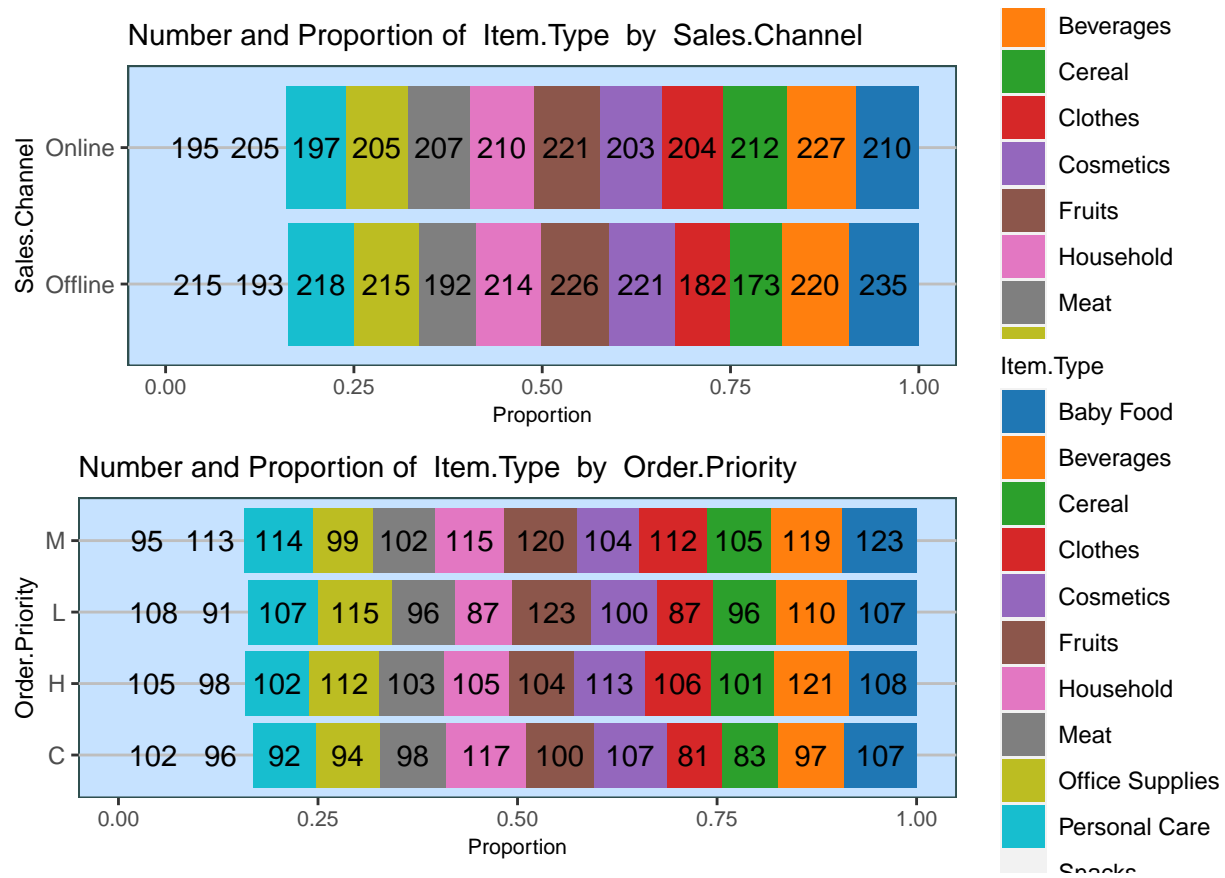


Item Type shows a similar lack of relationship to the other variables, though there is a bit more. But there is one major exception. Now we see the source of the singularity - each item type has one, and only one, unit price and vice versa. The two are completely correlated. Just to be sure, a regression shows an  $R^2$  of 1.

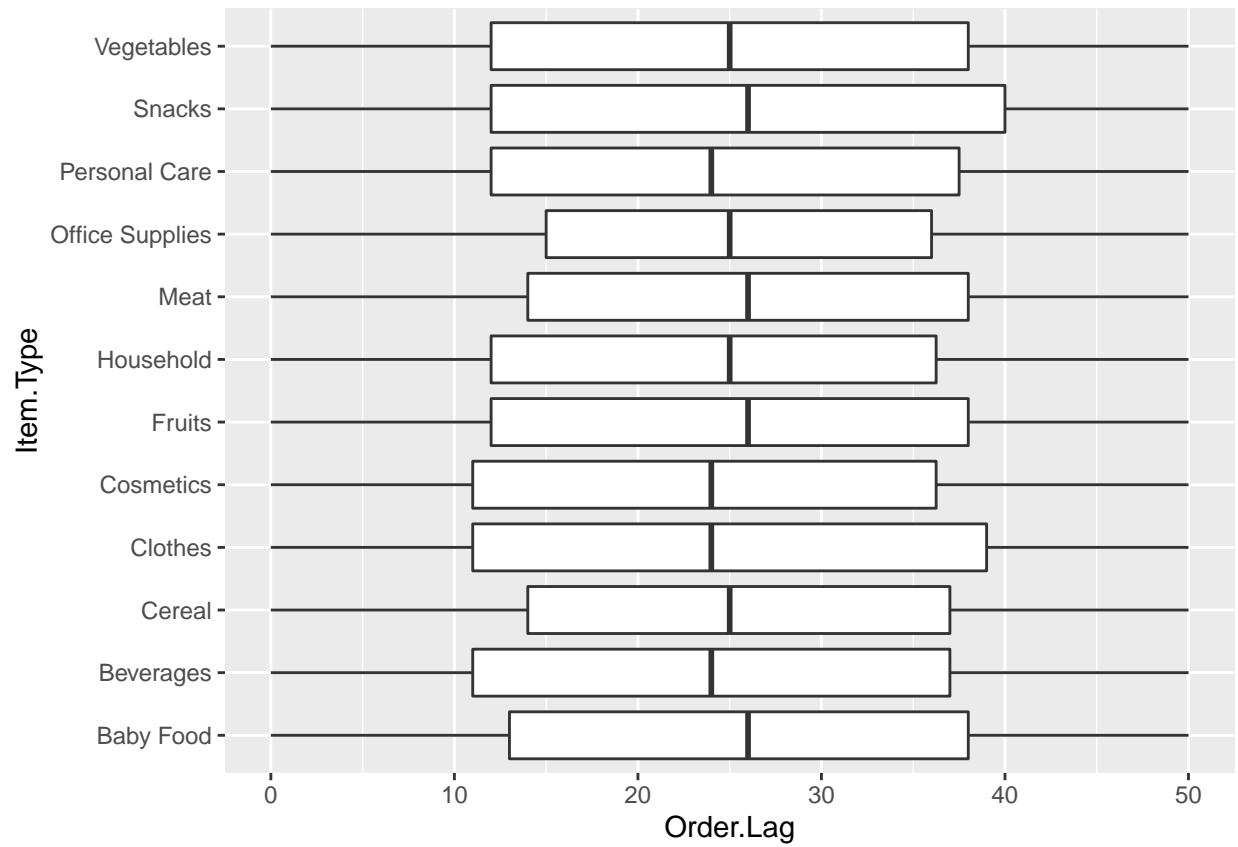
```
a <- EHExplore_TwoCategoricalColumns_Barcharts(df3, "Item.Type")
grid.arrange(grobs=a[c(3:4)])
```

```
## Warning: This manual palette can handle a maximum of 10 values. You have
## supplied 12.
```

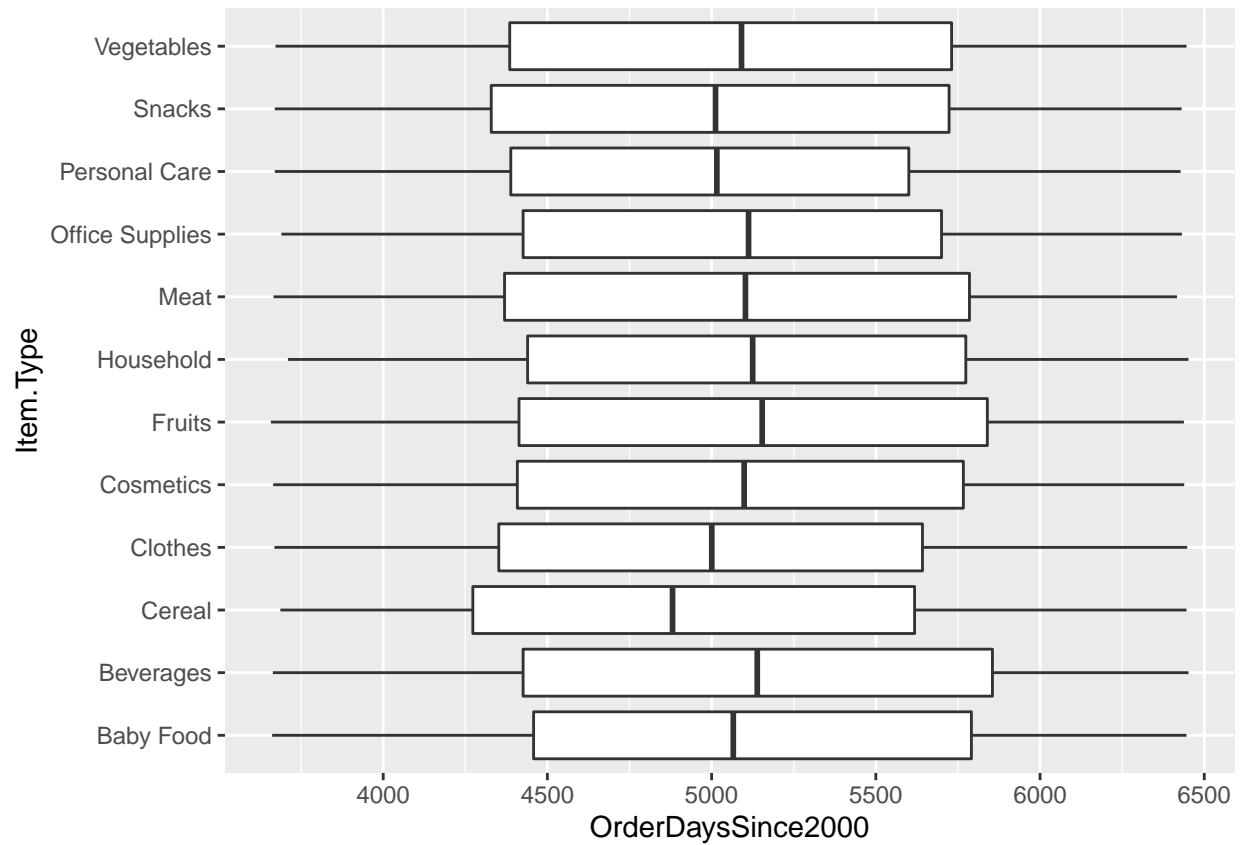
```
## Warning: This manual palette can handle a maximum of 10 values. You have
## supplied 12.
```



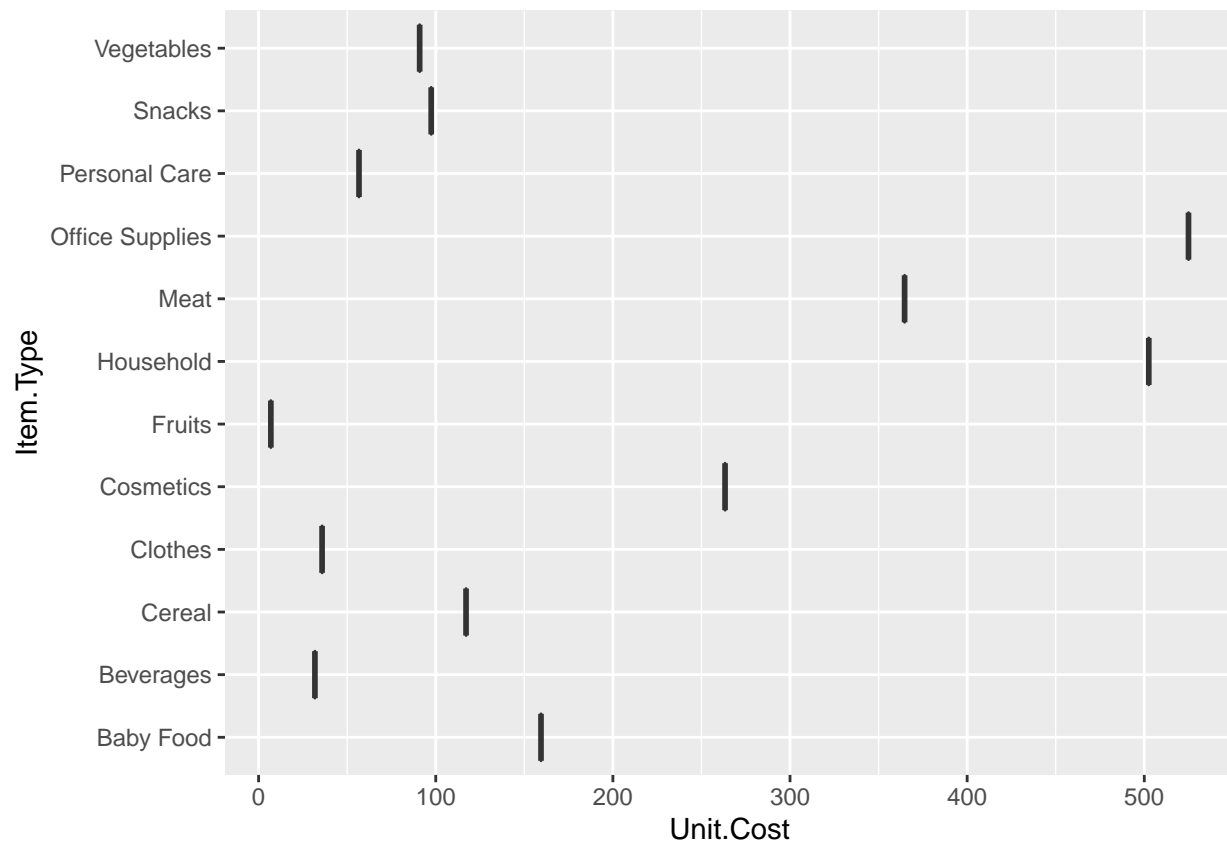
```
ggplot(df2, aes(Item.Type, Order.Lag)) +
  geom_boxplot() +
  coord_flip()
```



```
ggplot(df2, aes(Item.Type, OrderDaysSince2000)) +  
  geom_boxplot() +  
  coord_flip()
```



```
ggplot(df2, aes(Item.Type, Unit.Cost)) +  
  geom_boxplot() +  
  coord_flip()
```



```
dfx3 <- df3 %>%
  dplyr::select(Unit.Cost, Item.Type)

x <- lm(Unit.Cost ~ Item.Type, data=dfx3)
summary(x)
```

```
##
## Call:
## lm(formula = Unit.Cost ~ Item.Type, data = dfx3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.263e-10 -1.170e-13  0.000e+00  5.100e-14  2.305e-10
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)    1.594e+02  2.919e-13  5.461e+14 <2e-16 ***
## Item.TypeBeverages -1.276e+02  4.124e-13 -3.095e+14 <2e-16 ***
## Item.TypeCereal    -4.231e+01  4.286e-13 -9.872e+13 <2e-16 ***
## Item.TypeClothes   -1.236e+02  4.283e-13 -2.885e+14 <2e-16 ***
## Item.TypeCosmetics  1.039e+02  4.179e-13  2.486e+14 <2e-16 ***
## Item.TypeFruits    -1.525e+02  4.124e-13 -3.698e+14 <2e-16 ***
## Item.TypeHousehold  3.431e+02  4.179e-13  8.211e+14 <2e-16 ***
## Item.TypeMeat       2.053e+02  4.245e-13  4.835e+14 <2e-16 ***
## Item.TypeOffice Supplies 3.655e+02  4.189e-13  8.726e+14 <2e-16 ***
```

```
## Item.TypePersonal Care    -1.028e+02  4.202e-13 -2.445e+14   <2e-16 ***
## Item.TypeSnacks          -6.198e+01  4.248e-13 -1.459e+14   <2e-16 ***
## Item.TypeVegetables      -6.849e+01  4.215e-13 -1.625e+14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.158e-12 on 4988 degrees of freedom
## Multiple R-squared:      1, Adjusted R-squared:      1
## F-statistic: 3.73e+29 on 11 and 4988 DF, p-value: < 2.2e-16
```

With Unit Cost in the analysis, a machine learning exploration is not justified, since a lookup table in Excel would perform just as well. We will use total profit alone - this contains unit cost, but the correlation is low enough that the exercise is still worth doing. We remove Unit.Cost, dummify the categorical variables and scale all the predictors.

```
z=list("Item.Type")

df3a <- EHPrepare_CreateDummies(df3, exclude=z, dropFirst=TRUE)
```

```
## Warning: Predicate functions must be wrapped in 'where()'.
##
## # Bad
## data %>% select(is.factor)
##
## # Good
## data %>% select(where(is.factor))
##
## i Please update your code.
## This message is displayed once per session.
```

```
## Warning: Predicate functions must be wrapped in 'where()'.
##
## # Bad
## data %>% select(is.character)
##
## # Good
## data %>% select(where(is.character))
##
## i Please update your code.
## This message is displayed once per session.
```

```
## Warning in EHPrepare_CreateDummies(df3, exclude = z, dropFirst = TRUE): NAs
## introduced by coercion
```

```
df4 <- EHPrepare_ScaleAllButTarget(df3a, "Item.Type")
```

## B. Distributions