US Regional Sales Data Model

Arinze Francis

2022-06-10

```
## function (..., list = character(), pos = -1, envir = as.environment(pos),
##     inherits = FALSE)
## {
##     dots <- match.call(expand.dots = FALSE)$...
##     if (length(dots) && !all(vapply(dots, function(x) is.symbol(x) ||
##         is.character(x), NA, USE.NAMES = FALSE)))
##         stop("... must contain names or character strings")
##     names <- vapply(dots, as.character, "")
##     if (length(names) == 0L)
##         names <- character()
##     list <- .Primitive("c")(list, names)
##     .Internal(remove(list, envir, inherits))
## }
## <bytecode: 0x0000000014d25230>
## <environment: namespace:base>
```

Load R packages

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
## Warning: package 'equatiomatic' was built under R version 4.1.3
```

Import data

```
data <- read_xlsx("US_Regional_Sales_Data.xlsx")
data
```

```
## # A tibble: 7,991 x 16
##    OrderNumber `Sales Channel` WarehouseCode ProcuredDate
##    <chr>       <chr>           <chr>         <dttm>
##  1 SO - 000101 In-Store        WARE-UHY1004  2017-12-31 00:00:00
##  2 SO - 000102 Online          WARE-NMK1003  2017-12-31 00:00:00
##  3 SO - 000103 Distributor     WARE-UHY1004  2017-12-31 00:00:00
##  4 SO - 000104 Wholesale       WARE-NMK1003  2017-12-31 00:00:00
##  5 SO - 000105 Distributor     WARE-NMK1003  2018-04-10 00:00:00
##  6 SO - 000106 Online          WARE-PUJ1005  2017-12-31 00:00:00
##  7 SO - 000107 In-Store        WARE-XYS1001  2017-12-31 00:00:00
##  8 SO - 000108 In-Store        WARE-PUJ1005  2018-04-10 00:00:00
##  9 SO - 000109 In-Store        WARE-PUJ1005  2017-12-31 00:00:00
## 10 SO - 000110 In-Store        WARE-UHY1004  2017-12-31 00:00:00
## # ... with 7,981 more rows, and 12 more variables: OrderDate <dttm>,
## #   ShipDate <dttm>, DeliveryDate <dttm>, CurrencyCode <chr>,
## #   _SalesTeamID <dbl>, _CustomerID <dbl>, _StoreID <dbl>, _ProductID <dbl>,
## #   Order Quantity <dbl>, Discount Applied <dbl>, Unit Price <dbl>,
## #   Unit Cost <dbl>
```

Data Manipulations

```
any(is.na(data))
```

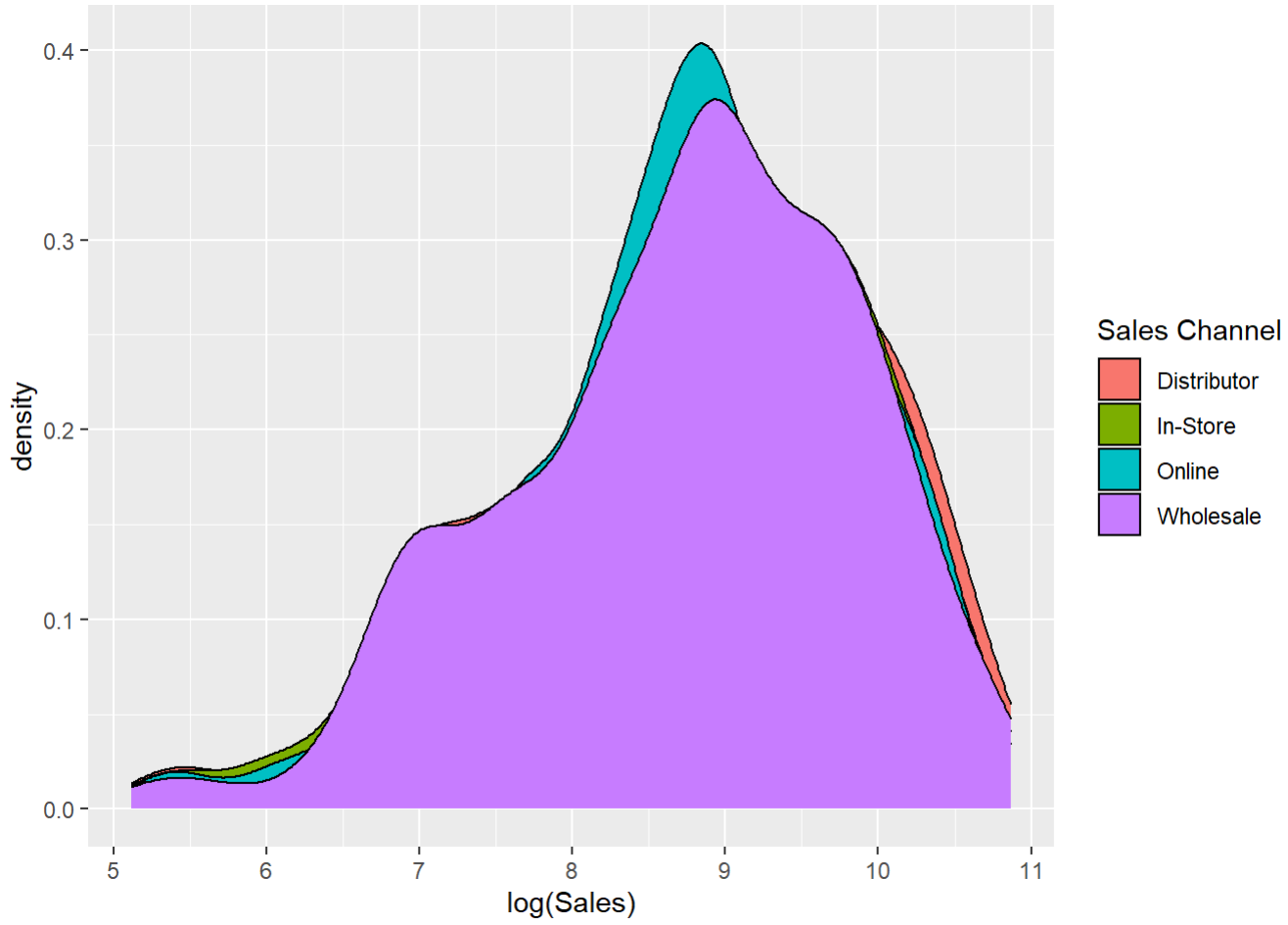
```
## [1] FALSE
```

```
data <- data %>% mutate(Sales = data$`Unit Price` * data$`Order Quantity`) # to get my sales Pri
ce

setnames(data, old = c('_ProductID','_StoreID'), new = c('Product_ID','Store_ID')) # Renaming of
relevant multiple columns

data_1 <- data %>% select(Sales, `Unit Price`, `Discount Applied`, `Sales Channel`, Product_ID,
 Store_ID ) # selection of relevant columns
```

Data Sales and Sales Channel Visualizaion

```
data %>% ggplot(aes(log(Sales), group = `Sales Channel`)) + geom_density(aes(fill=`Sales Channel
`),alpham= 0.8, colour = 'black')
```

```
## Warning: Ignoring unknown parameters: alpham
```



Estimate linear additive model

```
data_result <- lm(Sales ~ `Unit Price` + (`Discount Applied`+1) +`Sales Channel` , data = data_1
)

data_result
```

```
##
## Call:
## lm(formula = Sales ~ `Unit Price` + (`Discount Applied` + 1) +
##     `Sales Channel`, data = data_1)
##
## Coefficients:
##              (Intercept)              `Unit Price`        `Discount Applied`
##                   157.85                      4.54                    278.65
##  `Sales Channel`In-Store     `Sales Channel`Online  `Sales Channel`Wholesale
##                  -184.55                   -424.28                   -161.74
```

Multiplicative Model

```
data_result_11 <- lm(log(Sales) ~ log(`Unit Price`) + log(`Discount Applied` + 1) + `Sales Chann
el`, data = data_1 )
summary(data_result_11)
```

```
##
## Call:
## lm(formula = log(Sales) ~ log(`Unit Price`) + log(`Discount Applied` +
##     1) + `Sales Channel`, data = data_1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.352 -0.252  0.258  0.603  0.783
##
## Coefficients:
##                             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)                  1.33673    0.06309   21.19 <0.0000000000000002 ***
## log(`Unit Price`)            1.00128    0.00806  124.24 <0.0000000000000002 ***
## log(`Discount Applied` + 1) -0.08976    0.10306   -0.87                0.38
## `Sales Channel`In-Store     -0.00642    0.02126   -0.30                0.76
## `Sales Channel`Online       -0.01881    0.02236   -0.84                0.40
## `Sales Channel`Wholesale     0.00832    0.02846    0.29                0.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.662 on 7985 degrees of freedom
## Multiple R-squared:  0.659,  Adjusted R-squared:  0.659
## F-statistic: 3.09e+03 on 5 and 7985 DF,  p-value: <0.0000000000000002
```

Conclusions and Managerial Implications

66% of the fluctuations of sales can be explained by the model.Sales benefited greatly from the Wholesale sales channels by (0.008). In-store sales channel had the lowest contribution to sales by (0.006). If Wholesale sales channels is increased by one percent, the units sold is to increase by 0.00832 percent (holding all other variables constant) and etc.

Additional Insights

```
tidy_result<-tidy(data_result_11)
tidy_result
```

```
## # A tibble: 6 x 5
##   term                        estimate std.error statistic  p.value
##   <chr>                          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                   1.34      0.0631     21.2   5.43e-97
## 2 log(`Unit Price`)             1.00      0.00806   124.    0
## 3 log(`Discount Applied` + 1) -0.0898    0.103      -0.871 3.84e- 1
## 4 `Sales Channel`In-Store     -0.00642   0.0213     -0.302 7.63e- 1
## 5 `Sales Channel`Online       -0.0188    0.0224     -0.841 4.00e- 1
## 6 `Sales Channel`Wholesale     0.00832   0.0285      0.292 7.70e- 1
```

Model validation

Assumption 1: homoskedasticity

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:equatiomatic':
##
##     hsb
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
data_result_11 %>% ols_test_breusch_pagan()
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                  Data
##  -------------------------------------
##  Response : log(Sales)
##  Variables: fitted values of log(Sales)
##
##       Test Summary
##  -------------------------
##  DF            =    1
##  Chi2          =    0.1084
##  Prob > Chi2   =    0.7420
```

```
# Interpretation: the p-value is above an appropriate threshold (p < 0.05) therefore the null hy
pothesis of homoskedasticity is accepted
```

Assumption 2: MULTICOLLINEARITY

```
vif(data_result_11)
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## log(`Unit Price`)              1  1               1
## log(`Discount Applied` + 1)    1  1               1
## `Sales Channel`                1  3               1
```

```
# Interpretation: From the GVIF,there is no multivariate problem as the explanatory variables we
re below 5.
```