correlation analysis to improve marketing performance

ERICK@GURU

2024-03-06

How to use correlation analysis to improve marketing performance IN R Programming'

Why correlation?

It is important to understand what drives relationships. For example, if you want to determine how marketing performance impact your sales numbers you have to account for all factors that can explain your sales numbers. This could include factors like:

1. marketing channels
2. season
3. geographic location
4. etc.

Correlation is a great exploratory tool that can sometimes reveal interesting patterns in your data. Most importantly, once you get the hang of it, it is really easy to add to your analysis toolkit.

What is correlation?

A correlation is a statistic that quantifies the strength between 2 variables. The statistic is called the correlation coefficient denoted as r.

The correlation coefficient or r is a number between +1 and -1 and is calculated so as to represent the linear relationship between two variables. An r close to 1 indicates a strong relationship between the variables while an r close to 0 indicates a weak relationship.

A positive or negative sign indicates the direction of the relationship. A positive r indicates a positive relationship and a negative r a negative relationship. Also, we can plot the statistic in a correlation plot or matrix *(which we will do shortly)*.

Let's cover three (3) common correlation methods:

1. Pearson method — correlation is the default for linear relationships and assumes your data is normally distributed. It is sensitive to outliers and skewed data.
2. Spearman method — for non-normal populations. Checks for rank or ordered relationships.
3. Kendall method — for when you have a small dataset and many tied or rank relationships.

Your choice of correlation method should be driven by the underlying distribution of your data.

How to interpret correlation

Correlation thresholds using Jacob Cohen's Rule of Thumb which is often used in the behavioral sciences to interpret the effect size:

r >= 0.5 large or strong association
r = 0.3 medium association
r = 0.1 small or weak association

If the underlying data distribution is not normal, then you could transform (e.g. logarithm, Box-Cox, etc.) your variables before attempting to apply these thresholds.

What's an acceptable correlation?

Even if the correlation coefficient is **at** or **near** zero, that doesn't mean no relationship exists. It's just that relationship isn't linear, but there could be other relationships which is why it's important to visualize your variables beforehand.

Exploratory Visualization and Correlation Analysis

We will be using this marketing dataset that is available on Kaggle.

The data contains the sales data for two consecutive years of a product of a non-specified brand. Each row contains the Volume of Sales for a week and includes additional information or various promotion methods for that product for each week. Let's inspect the dataset.

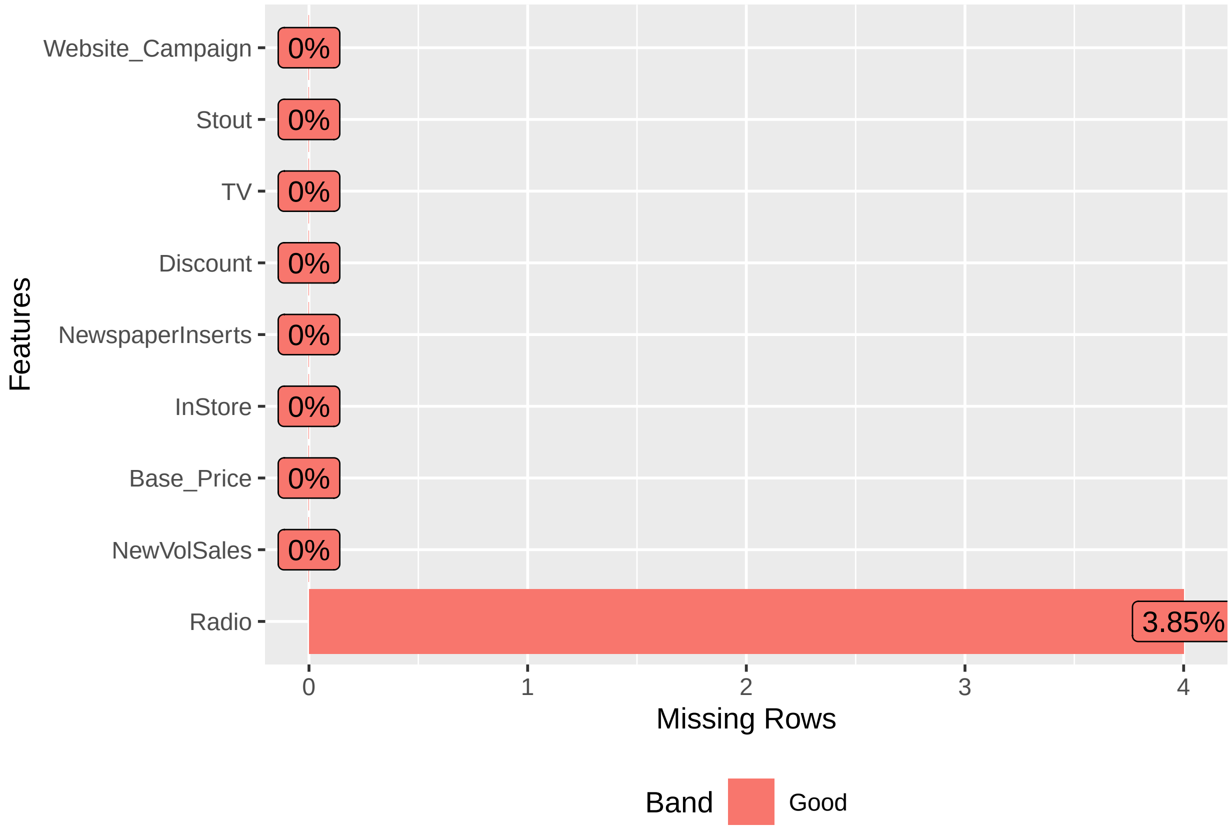
```r
#libraries
library(tidyverse)
library(dplyr)
library(corrplot)
library(VIM)
library(cowplot)
library(parsnip)
library(tidymodels)
library(DataExplorer)
```

```r
#load data
df <- read.csv("C:/Users/langa/OneDrive/Desktop/R PROGRAMMING PRACTICE/Learning Labs/correlation analysi
#str
str(df)
```

```
## ’data.frame’:    104 obs. of  9 variables:
##  $ NewVolSales     : int  19564 19387 23889 20055 20064 19481 19509 19033 20498 19509 ...
##  $ Base_Price      : num  15 15 14.6 15.3 15.6 ...
##  $ Radio           : num  245 314 324 298 279 259 235 290 315 318 ...
##  $ InStore         : num  15.5 16.4 62.7 16.6 41.5 ...
##  $ NewspaperInserts: chr  "" "" "" "" ...
##  $ Discount        : num  0 0 0.05 0 0.045 0 0 0 0.035 0.045 ...
##  $ TV              : num  101.8 76.7 131.6 119.6 103.4 ...
##  $ Stout           : num  2.28 2.22 2.01 2.2 1.82 ...
##  $ Website_Campaign: chr  "" "" "" "" ...
```

Exploratory data visualization

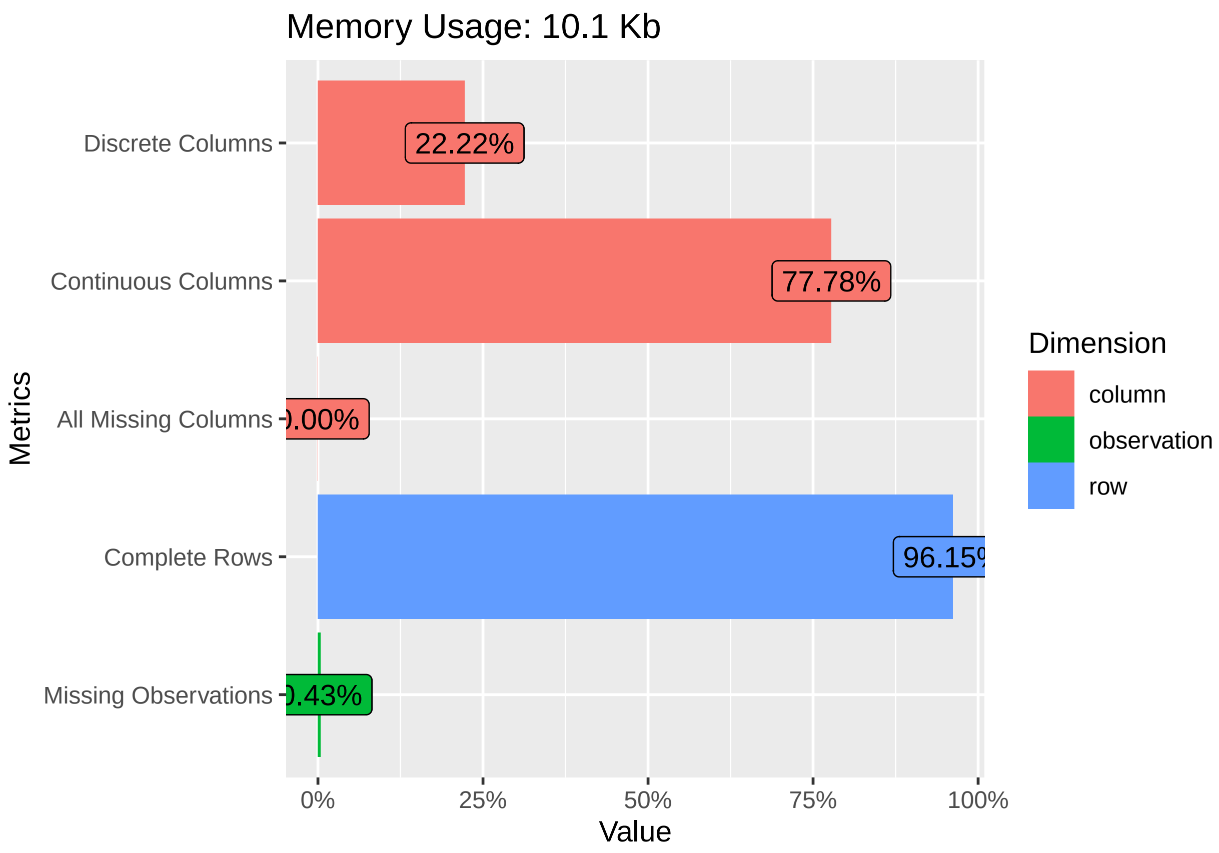
```
#EDA
df %>%
    plot_missing()
```



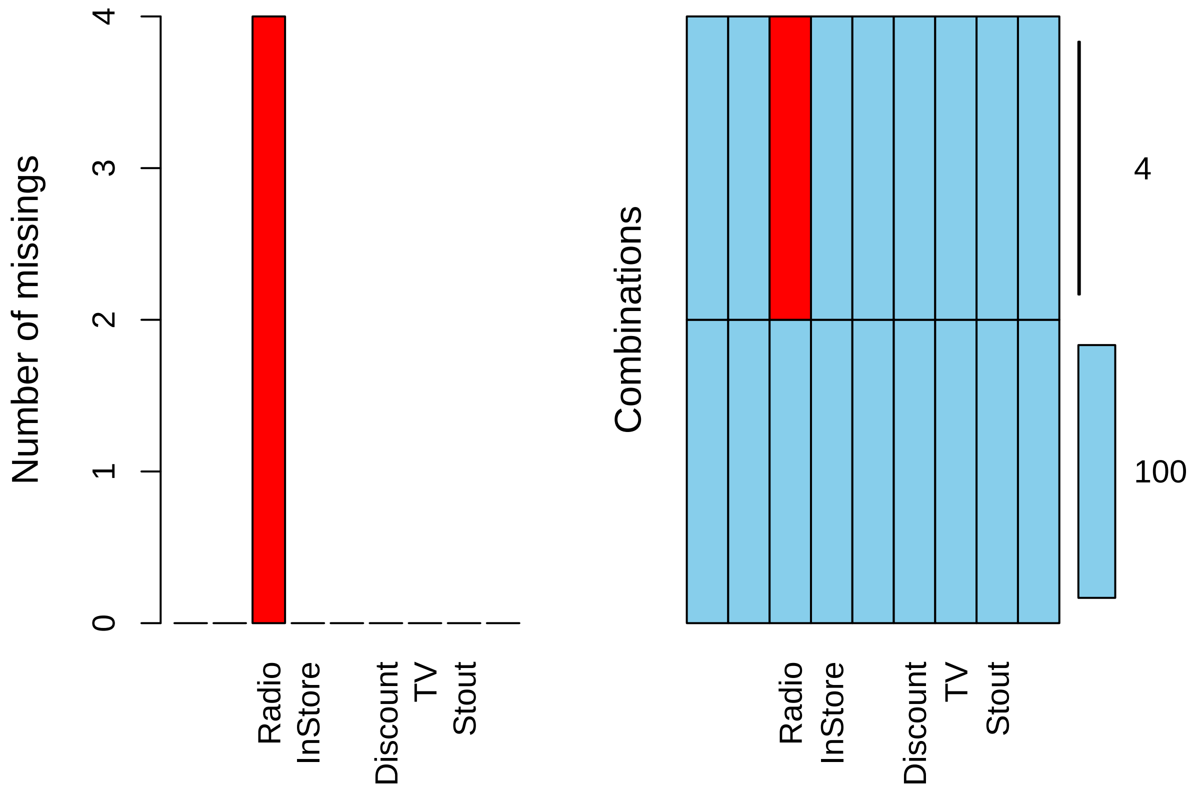
```
df %>% profile_missing()
```

```
##             feature num_missing pct_missing
## 1       NewVolSales           0  0.00000000
## 2        Base_Price           0  0.00000000
## 3             Radio           4  0.03846154
## 4           InStore           0  0.00000000
## 5 NewspaperInserts           0  0.00000000
## 6          Discount           0  0.00000000
## 7                TV           0  0.00000000
## 8             Stout           0  0.00000000
## 9 Website_Campaign           0  0.00000000
```

```
df %>%  plot_intro()
```



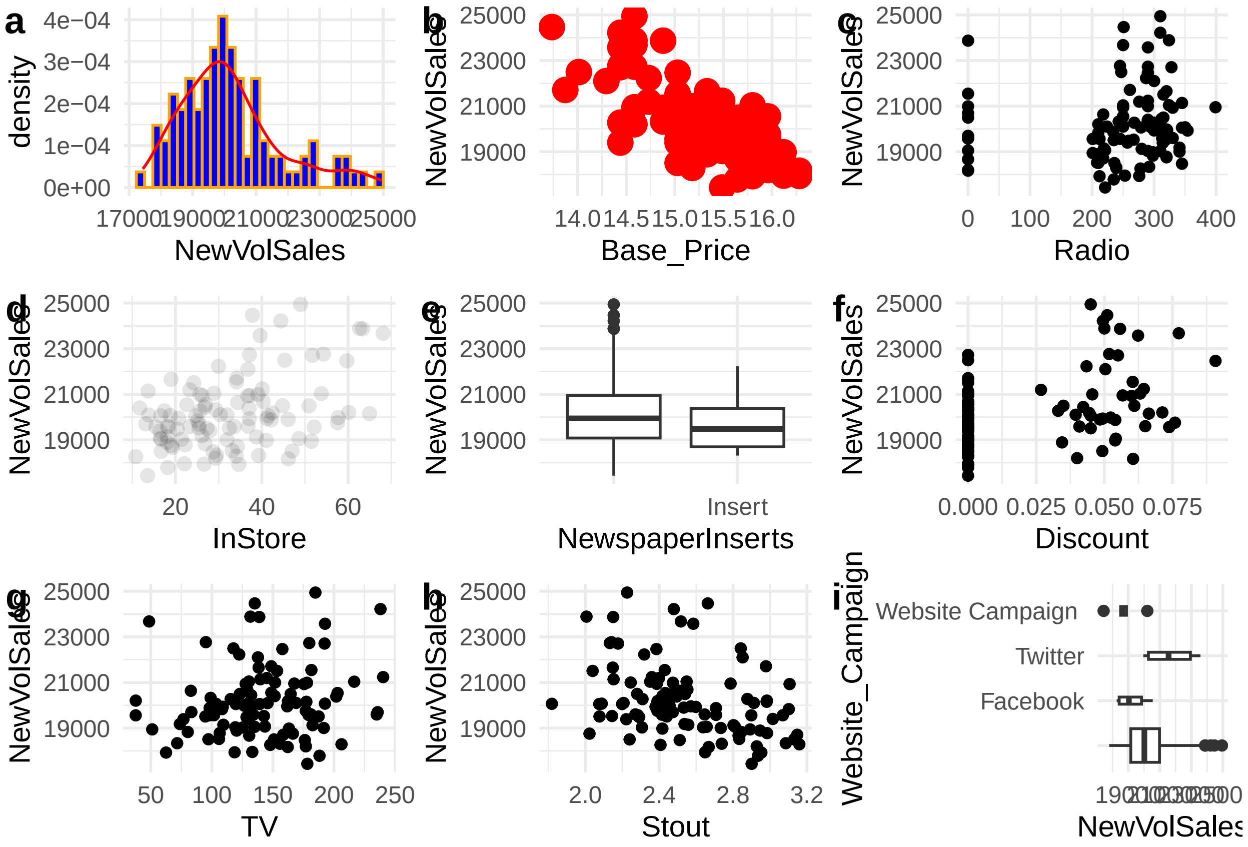
```
#any missing data? Using VIM package
aggr(df, prop = F, numbers = T) # radio has some missing data
```



```r
#We will place with a 0# replace NA'S with Zero
df$Radio[is.na(df$Radio)] <- 0
```

```r
p1 <- df %>% ggplot(aes(NewVolSales)) +
      geom_histogram(bins = 30, aes(y=..density..),
      colour="orange", fill="blue") +
     geom_density(alpha=0.1, color='red') + theme_minimal()

p2 <- df %>%  ggplot( aes( Base_Price, NewVolSales)) +
              geom_point(size= 4, color='red') + theme_minimal()


p3 <- ggplot(df, aes(Radio, NewVolSales)) + geom_point() + theme_minimal()

p4 <- ggplot(df, aes(InStore, NewVolSales)) + geom_point(size=2, alpha=1/10)     + theme_minimal()
p5 <- ggplot(df, aes(NewspaperInserts, NewVolSales)) + geom_boxplot() +            theme_minimal()
p6 <- ggplot(df, aes(Discount, NewVolSales)) + geom_point() +
        theme_minimal()
p7 <- ggplot(df, aes(TV, NewVolSales)) + geom_point() + theme_minimal()

p8 <- ggplot(df, aes(Stout, NewVolSales)) + geom_point() + theme_minimal()

p9 <- ggplot(df, aes(Website_Campaign, NewVolSales)) +
        geom_boxplot(varwidth = TRUE) + theme_minimal() + coord_flip()

plot_grid(p1, p2, p3, p4, p5, p6, p7, p8, p9, labels = "auto")
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

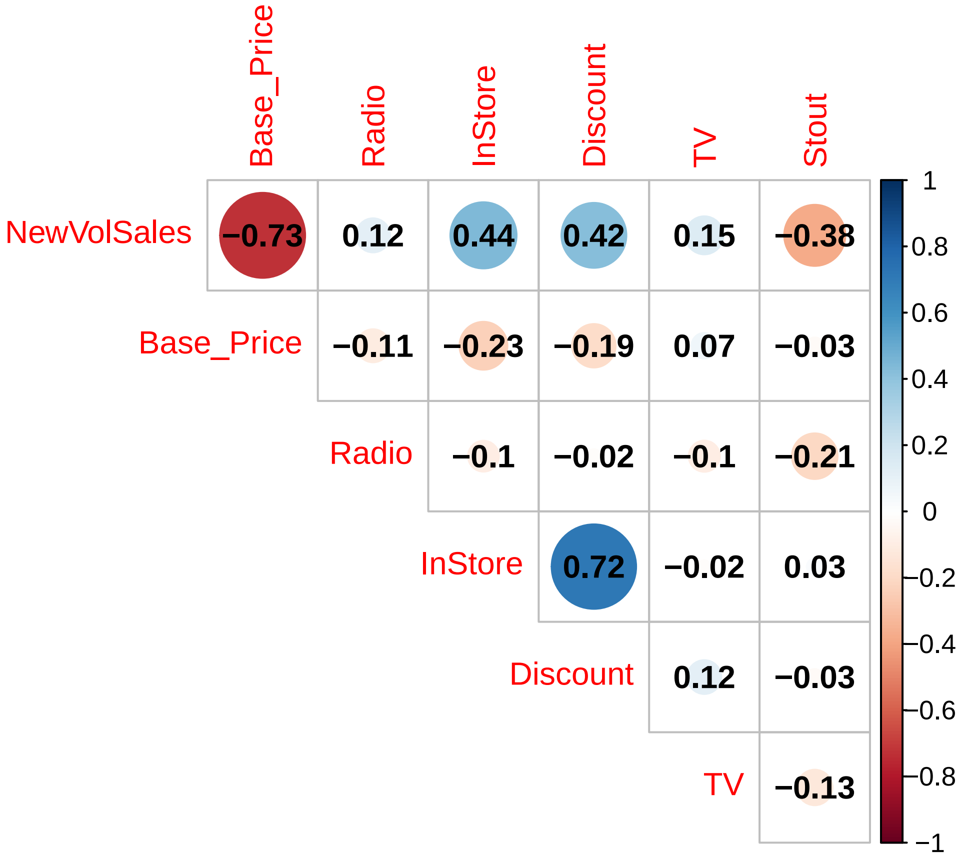


What can we tell from the exploratory data visualization?

- Sales appear to be relatively normal with perhaps a bit of right skew, but not enough to be particularly worrying.
- Higher sales when the base price is low. Less sales when base price is high.
- Stout seems to have a negative impact on sales. Not sure what stout refers to as it was not included in the data dictionary and the data source is anonymous.
- InStore appears to have a positive impact on sales.
- Both Radio & TV impact inconclusive. More sales when radio & tv spending is going on, but not necessarily always the case.
- Newspaper insert doesn't appear to have any significant impact on sales.
- Website Campaign appears to have more sales when there is Twitter engagement, but it doesn't appear to be significantly different when there is no website campaign going on.

Next, let's create our correlation plot.

```
#R
df %>%
 select_if(is.numeric) %>%
  cor() %>%
    corrplot(type = "upper", addCoef.col = "black", diag=FALSE)
```



We have the correlation coefficients in each box. Positive correlations are in blue. Negative correlations are in red.

Summary of correlations:

- Instore and discount both have a medium positive correlation to NewVolSales.
- Radio and TV have a weak positive correlation to NewVolSales.
- Price has a strong negative correlation to NewVolSales.
- Last, but not least, Stout has a medium negative correlation to NewVolSales

> **NOTE**: The coefficient of determination is our correlation coefficient squared. It is the proportion of the variance in the y (dependent) variable that is predictable from our x (independent) variable.

Correlation and it's relationship to regression

Let's review how correlation and regression are related by reviewing just 2 variables (NewVolSales and Discount).

The correlation coefficient of NewVolSales and Discount ads is 0.42 (rounded to 2 decimal places).

```
#R
round(cor(df$NewVolSales, df$Discount),2)
```

```
## [1] 0.42
```

```
0.42
```

```
## [1] 0.42
```

If we model this in a linear regression model and extract the r-squared, the result is 0.18. (rounded to 2 decimal places).

```
# R
linear_model <- lm(NewVolSales ~ Discount, data = df)
summary(linear_model)
```

```
##
## Call:
## lm(formula = NewVolSales ~ Discount, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2934.5  -986.7  -144.2   751.3  4217.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19636.5      180.4 108.826  < 2e-16 ***
## Discount     24234.8     5114.4   4.739 6.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1436 on 102 degrees of freedom
## Multiple R-squared:  0.1804, Adjusted R-squared:  0.1724
## F-statistic: 22.45 on 1 and 102 DF,  p-value: 6.987e-06
```

If we square the correlation coefficient of 0.42, we will get our r-squared 0.180. In effect, the correlation coefficient squared is the r-squared.

Testing for Significance

Let's test significance of all the variables in our dataset by using a linear regression model on the entire dataset.

```r
# R (Full Model)
model_spec_lm <- linear_reg() %>%
    set_engine('lm') %>%
    set_mode('regression')
mkmix_model <- model_spec_lm %>%
    fit(NewVolSales ~ Base_Price + Radio + InStore + factor(NewspaperInserts) + Discount + TV + Stout +
summary(mkmix_model$fit)
```

```
##
## Call:
## stats::lm(formula = NewVolSales ~ Base_Price + Radio + InStore +
##     factor(NewspaperInserts) + Discount + TV + Stout + factor(Website_Campaign),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1523.13  -473.88    16.83   381.71  2018.65
##
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                 54394.3976  2278.7953  23.870
## Base_Price                                  -2056.8098   136.7813 -15.037
## Radio                                          -0.2801     0.7214  -0.388
## InStore                                        28.6341     7.3836   3.878
## factor(NewspaperInserts)Insert                159.0085   289.0542   0.550
## Discount                                     5554.2422  3618.1645   1.535
## TV                                              3.2233     1.6875   1.910
## Stout                                       -1631.4978   235.9605  -6.914
## factor(Website_Campaign)Facebook              239.8193   362.6601   0.661
## factor(Website_Campaign)Twitter               468.7707   362.9915   1.291
## factor(Website_Campaign)Website Campaign  -1273.3869   315.4384  -4.037
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## Base_Price                                   < 2e-16 ***
## Radio                                       0.698694
## InStore                                     0.000196 ***
## factor(NewspaperInserts)Insert              0.583570
## Discount                                    0.128155
## TV                                          0.059199 .
## Stout                                       5.83e-10 ***
## factor(Website_Campaign)Facebook            0.510069
## factor(Website_Campaign)Twitter             0.199762
## factor(Website_Campaign)Website Campaign  0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 677.4 on 93 degrees of freedom
## Multiple R-squared:  0.8337, Adjusted R-squared:  0.8159
## F-statistic: 46.63 on 10 and 93 DF,  p-value: < 2.2e-16
```

Our baseline revenue is 54,394. The only positive significant *(p-value is less than 0.05)* variable is InStore. The negative significant variables are Base_Price, Stout & Website Campaign *(no campaign)*.

The generic interpretation for each of our coefficients is for every one unit increase in the x variable, the y variable *(NewVolSales)* increases by beta units.

For example, for every 1 unit increase of InStore, sales increase by 28. If our sales volume is in dollars, then this would be a 28 dollar increase.

What can we derive from our correlation analysis and how can we use this to inform marketing?

We will just focus on the relationship between ***NewVolSales*** and each independent (x) variable.

What's **not working** in marketing?

1. Price — *significant.* We lose money when we increase the base price.
2. Stout — *significant*
3. Website Campaign — *significant.* This means when there is no website campaign going on. This has a negative effect on sales.

What is **working** with marketing?

1. Instore — *significant*

What is **not as impactful**?

1. Radio — *not significant*
2. TV — *not significant*
3. Discount — *not significant*

You should at this point have a conversation with your marketing stakeholders to understand their marketing goals and tactics for each of their marketing initiatives. There may be different goals for different initiatives. For instance, if marketing is using Radio and TV for top-funnel activities, then what we see in the data makes sense. Radio and TV are great for branding (e.g. awareness), but may have less of an impact on bottom-funnel metrics like sales. So those marketing efforts that aren't necessarily significant from a statistical point of view, can have impact from a real-world point of view.

Summary

With our correlation analysis we have derived some key insights into what is working and what isn't working when it comes to increased sales. Now, it's up to you to go further in your analysis which could include:

1. Adding in additional factors that were not included initially
2. Quantifying the impact of each marketing effort (e.g. ROI). In other words, calculate the marketing contribution towards sales.
3. Build a future forecast based on current levels of marketing spend and promotions.