Marketing Budget analysis and visualization

Hamza Belhaj

2023-04-18

Summary

In this R Markdown document, we will describe the process we used to analyze this marketing budget dataset in order to detect trends and identify how does the amount of each budget affect the sales.

Analysis

Loading Packages

The first step in our process is to load the different packages we will need into R. We use pacman to load the other packages because it does automatically install them too if they're messing, which is more convenient. We also set the directory we will be working with using the setwd() function.

```
pacman::p_load(pacman, caret, lars, tidyverse, psych, lmtest)
setwd("C:/Users/Hamza/Desktop/Projects/Marketing Budget")
getwd()
```

```
## [1] "C:/Users/Hamza/Desktop/Projects/Marketing Budget"
```

Loading data

Afterwards, we load and open our data

```
budget <- read.csv("Dummy Data HSS Clean.csv")
head(budget)
```

```
##   TV     Radio Social_Media Influencer     Sales
## 1 16  6.566231     2.907983       Mega  54.73276
## 2 13  9.237765     2.409567       Mega  46.67790
## 3 41 15.886446     2.913410       Mega 150.17783
## 4 83 30.020028     6.922304       Mega 298.24634
## 5 15  8.437408     1.405998      Micro  56.59418
## 6 29  9.614382     1.027163       Mega 105.88915
```

The describe function helps up get a general idea about our data by calculating some descriptive measurements. Note that Influencer is a categorical variable,

```
describe(budget)
```

```
##              vars    n   mean    sd median trimmed    mad  min    max  range
## TV              1 4546  54.06 26.10  53.00   53.82  32.62 10.0 100.00  90.00
## Radio           2 4546  18.16  9.66  17.86   17.99  11.11  0.0  48.87  48.87
## Social_Media    3 4546   3.32  2.21   3.06    3.16   2.39  0.0  13.98  13.98
## Influencer*     4 4546   2.51  1.11   3.00    2.51   1.48  1.0   4.00   3.00
## Sales           5 4546 192.41 93.02 188.96  191.52 118.41 31.2 364.08 332.88
##                skew kurtosis   se
## TV             0.07    -1.19 0.39
## Radio          0.14    -0.82 0.14
## Social_Media   0.65     0.05 0.03
## Influencer*  -0.01    -1.35 0.02
## Sales          0.07    -1.18 1.38
```

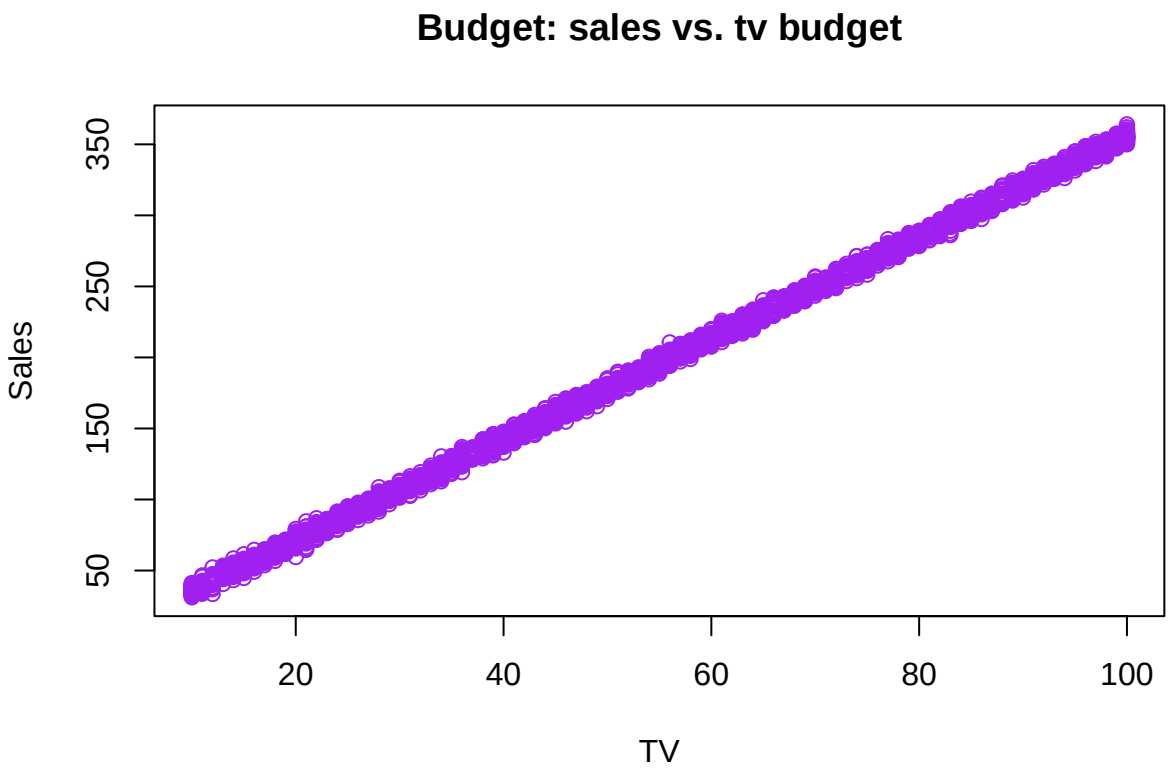
Some of the conclusions we can draw from this data:

- The mean and median values aren't very far apart for all variables, this indicates that there aren't too many outliers in our data.
- The different variables have varying degrees of standard deviation and range, with social media being the lowest, which means that for social media the values are more crowded near the mean and not as dispersed as the other variables.

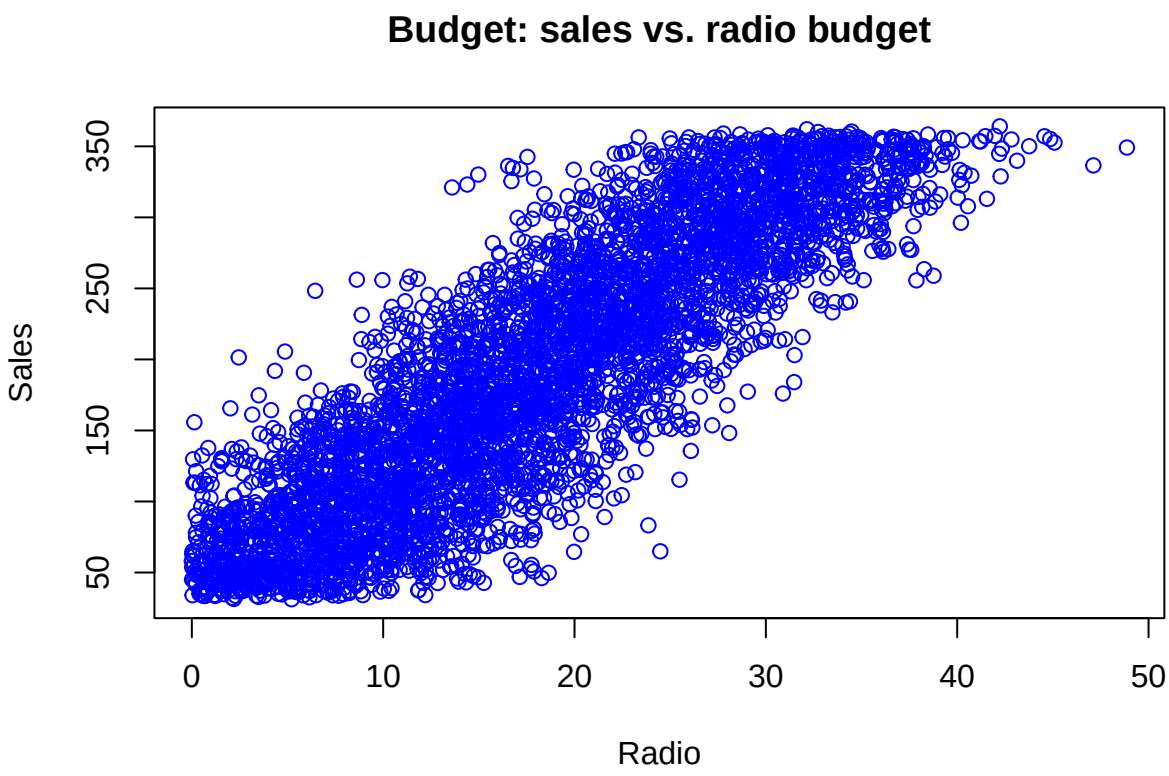
Determine the correlation between the sales and the different budgets

For this part, we will use 2 methods to verify correlation. First, we will plot the data to visually detect any trends in the data. Then we will calculate the correlation coefficient to be more precise.

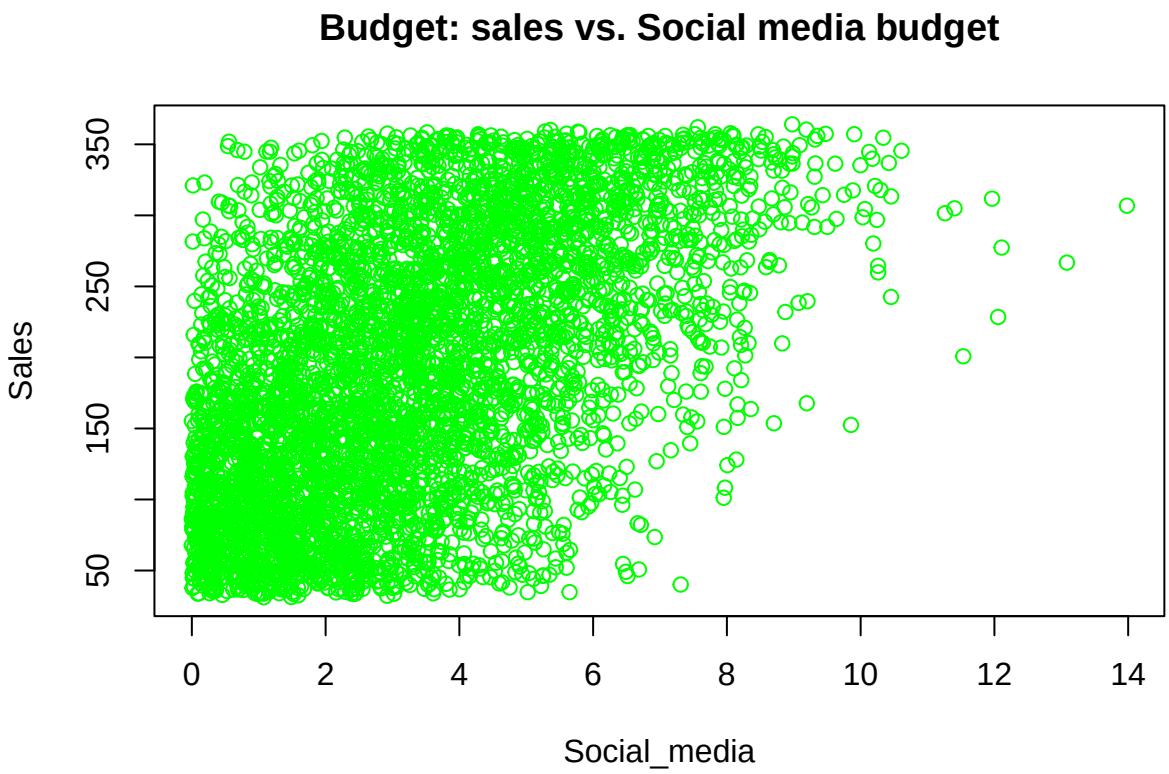
```
plot(budget$TV, budget$Sales,
     col = "purple",
     main = "Budget: sales vs. tv budget",
     xlab = "TV",
     ylab = "Sales")
```




```
plot(budget$Radio, budget$Sales,
     col = "blue",
     main = "Budget: sales vs. radio budget",
     xlab = "Radio",
     ylab = "Sales")
```




```
plot(budget$Social_Media, budget$Sales,
     col = "green",
     main = "Budget: sales vs. Social media budget",
     xlab = "Social_media",
     ylab = "Sales")
```



In these plots, we can already notice that there is a very strong correlation between TV budget and sales. There also some correlation between Radio budget and sales. While when it comes to social media, the correlation is less apparent.

Now we try to calculate the coefficient of correlation between sales and the different variables.

```
budget %>% summarise(cor(Sales,TV),cor(Sales,Radio),cor(Sales,Social_Media))
```

```
##   cor(Sales, TV) cor(Sales, Radio) cor(Sales, Social_Media)
## 1      0.9994974         0.8686378                0.5274464
```

The values of the coefficients confirm our observations from the plots.

Multiple Linear Regression

In the next step, we try to build a regression model that tells us the contribution of each budget to the sales. The sales are considered the dependent variable, while TV, Radio, And Social_media are our independent variables.

First of all, we start by building the model using the lm function.

```
reg1 <- lm(budget$Sales ~ budget$TV + budget$Radio + budget$Social_Media)
# Inferential tests
summary(reg1)
```

```
##
## Call:
## lm(formula = budget$Sales ~ budget$TV + budget$Radio + budget$Social_Media)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6158  -2.0002  -0.0075   2.0199  11.2581
##
## Coefficients:
##                       Estimate Std. Error  t value Pr(>|t|)
## (Intercept)         -0.133963   0.102820   -1.303    0.193
## budget$TV            3.562570   0.003389 1051.118   <2e-16 ***
## budget$Radio        -0.003970   0.009781   -0.406    0.685
## budget$Social_Media  0.004964   0.024884    0.199    0.842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.95 on 4542 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.999
## F-statistic: 1.505e+06 on 3 and 4542 DF,  p-value: < 2.2e-16
```

Based on the results of the inferential tests, we can conclude the following:

- The R-squared values are very close to 1, which means that the variance of the dependent variable can be explained by the dependent variables.
- The standard error is of 2.95. This means that the regression model predicts the sales with an average error of about 2.95.
- The p-value is less than 0.05. We can reject the null hypothesis and say that our module is statistically significant.
- From the coefficients for each variable, we can say that the TV variable is the most influential on sales.
- For the p-values of the coefficients, only the one for TV is less than 0.05. This means that we cannot reject the null hypothesis for the radio and social media budgets, and therefore we cannot conclude causation, even if we have high correlation. Because of this, it is also possible to create another model that predicts the value of sales based only on TV budget.

Single linear regression

Now we will create another regression model using only TV as our independent variable and compare it to the previous one.

```
reg2 <- lm(budget$Sales ~ budget$TV)
summary(reg2)
```

```
##
## Call:
## lm(formula = budget$Sales ~ budget$TV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6062  -2.0062  -0.0125   2.0249  11.2566
##
## Coefficients:
```

```
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -0.132493   0.100605   -1.317    0.188
## budget$TV    3.561514   0.001676 2125.272   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.949 on 4544 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.999
## F-statistic: 4.517e+06 on 1 and 4544 DF,  p-value: < 2.2e-16
```

By comparing the 2 models, we notice that The standard error is slightly less in the second model, the R-squared values are identical and the p-value is still less than 0.05. So it is easier to use the second model over the first one.

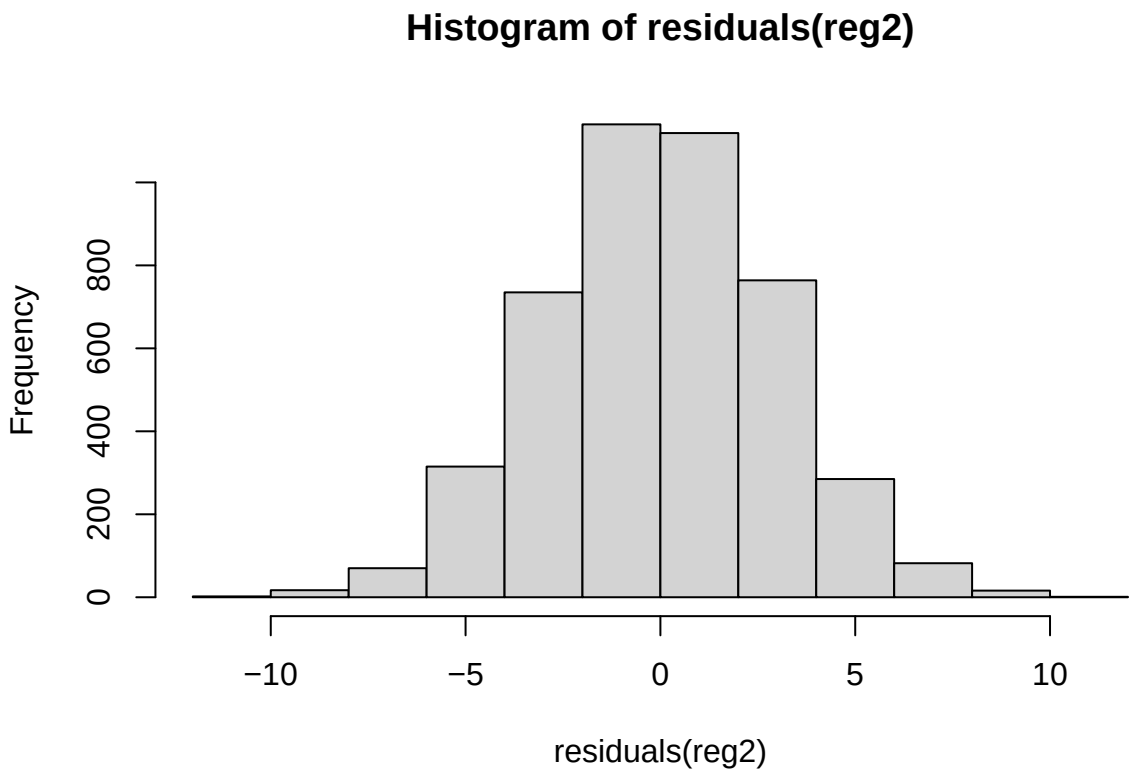
Verifying assumptions

In order for our interpretations of the results from the model to be significant, there are a number of assumptions that we need to check.

Linear relationship between dependent and independent variable This one can be easily verified using the plot between TV budget and sales. And as we can see, there is a very strong linear relationship between the 2.

Normally distributed error component epsilon This means that the error needs to follow a normal distribution. We can verify this using a histogram of the residuals

```
hist(residuals(reg2))
```



As we can see in the following histogram, the residuals do follow a normal distribution.

Absence of Heteroscedasticity In order for our model to be significant, the variance of the residuals needs to be the same across all values of the predicted variable, which is called homoscedasticity. One of the ways to verify for this is by performing the Breusch–Pagan Test.

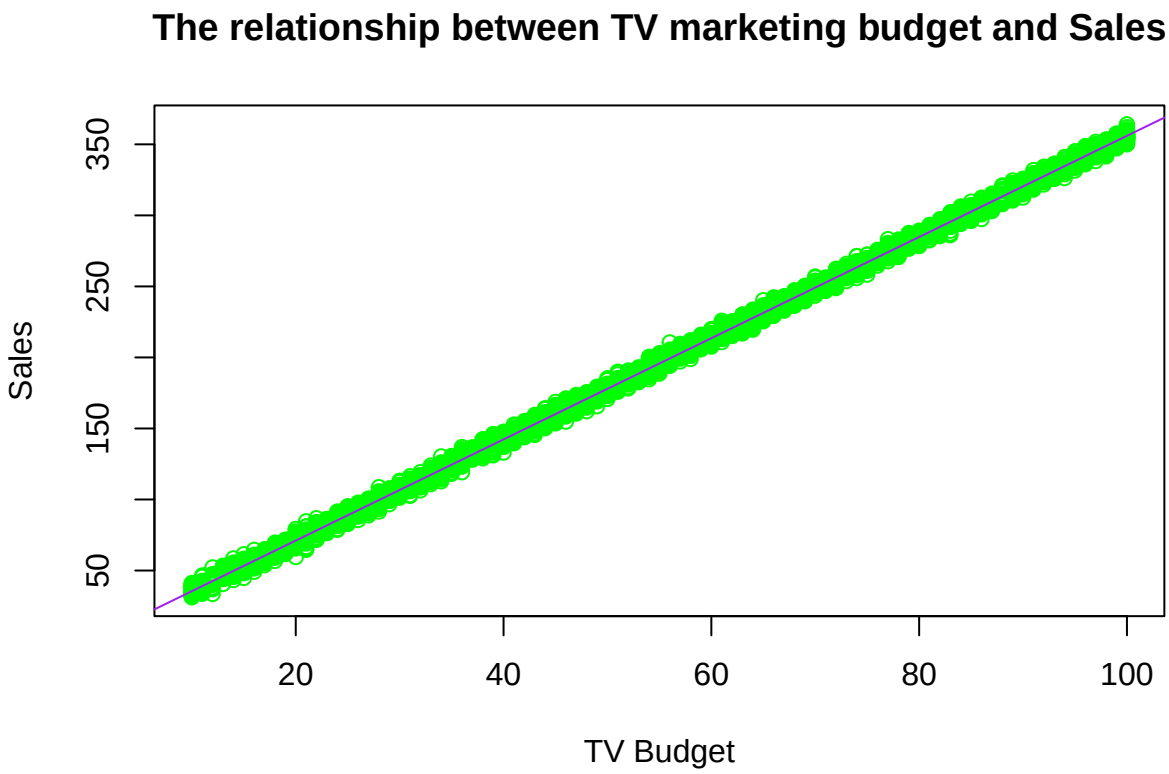
```
lmtest::bptest(reg2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg2
## BP = 5.9269e-05, df = 1, p-value = 0.9939
```

According to the Breusch-Pagan test, the BP value is small, and the p-value is over 0.05. Therefore, we do not reject the null hypothesis, and we assume that the residuals are homoscedastic.

Since our model passed all these tests, we can say with confidence that our results are statistically significant. Now all we have left is to plot the model.

```
plot(x=budget$TV, y=budget$Sales,col = 'green', main='The relationship between TV marketing budget and S
abline(reg2,col='purple')
```



Based on this model, we can conclude that for the current customer segments, TV is the optimal way to promote our products, since the more we invest into this communication channel, the bigger our sales are, and we established causation from this correlation.

These results make us ask the following questions:

- Why don't the other communication channels like radio and social media as impact on sales as TV does?
- Is it related to the nature of each channel? The type of messages used in each of them? Or is it related to the customer segments we are targeting?
- Can that be changed?