Report #0001.2

Advertising

Forni Lucas

May 1st, 2023

Executive Summary

This report presents an analysis of the relationship between advertising expenditure and sales in a business, obtained from a dataset provided by the company. Three advertising media were examined: television, radio, and newspaper. The results show that television advertising spending has the greatest positive impact on sales, followed by radio advertising spending. Newspaper advertising spending has a positive impact on sales but is weaker and not statistically significant when controlling for the other two media.

Key recommendations for the client are:

Prioritize spending on television and radio advertising, as they have the greatest positive impact on sales.

Reduce spending on newspaper advertising, as its impact on sales is less significant.

Monitor the effectiveness of advertising in each medium and adjust budgets accordingly.

Consider possible future research on synergy between advertising media and explore the possibility of using them in combination to maximize sales impact.

Futher experimentation with different metrics of performance.

Table of Contents

Content

Introduction

This project deals with the analysis of the relationship between advertising expenditures in several media and the sales of a product. The main objective of the analysis is to determine which advertising media have the greatest impact on product sales and how they relate to each other. In addition.

The scope of this analysis is limited to the data provided in the "Advertising.csv" dataset, which contains information on advertising expenditures in three different media (TV, radio and newspaper) and the corresponding product sales. It is assumed that the data are representative of the market and that there are no other important factors affecting the sales of the product other than the advertising spend in the three mentioned media. Limitations of this analysis include the lack of information on other factors that could influence product sales, as well as the limitation of the data set to only three advertising media.

The results reveal that spending on TV advertising has the largest positive impact on sales, followed by spending on radio advertising. Spending on newspaper advertising, while positively correlated with sales, is not statistically significant when controlling for the other two media.

The structure of the project will consist of an introduction, a discussion, where you will find a description of the data and data preparation, a multiple regression analysis to determine the relationship between advertising expenditures and product sales, an analysis of the correlation between the different advertising media. It will conclude with a conclusion of the results and possible implications for the advertising strategies of the product in question.

Discussion

Data understanding

The data used in this project is taken from a file called "advertising.csv", which is stored in a Google Drive folder provided by the teacher.

This data file provides information on advertising expenditures and corresponding sales for three types of media: TV, radio and newspapers. The dataset is composed of 200 observations and 5 variables, namely:

-ID: unique identifier of each observation.

-TV: TV advertising expenditure for each observation.

-Radio: radio advertising expenditure for each observation.

-Newspaper: newspaper advertising expenditure for each observation.

-Sales: sales for each observation.

The sales variable is in thousands of units, and the TV, radio and newspaper variables are in thousands of dollars.

For the followings histograms the number of bins was determined using the "Square-root rule", which suggests taking the square root of the total number of observations and rounding it to the nearest integer. In this case 14 or 15, 15 was the election.

TV

Name	Type	Missing	Min	Max	Average	Deviation
TV	Real	1	0.7	296.4	146.728	85.955

Table 1- TV

From this table:

The variable is a continuous numeric variable represented as a real number.

There is one missing value in the TV advertising variable.

The minimum value for TV advertising is $700, and the maximum value is $296400.

The average or mean of TV advertising expenditure is $146728.

The standard deviation of the TV advertising variable is $85955, which indicates a relatively large amount of variability in the data.

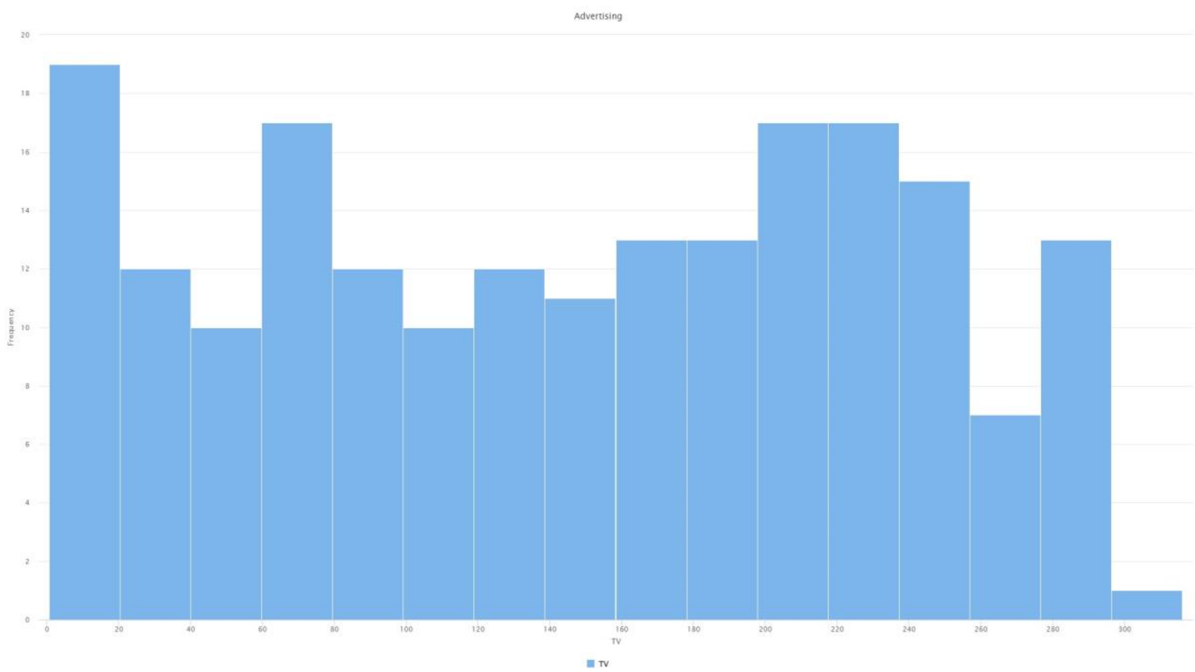
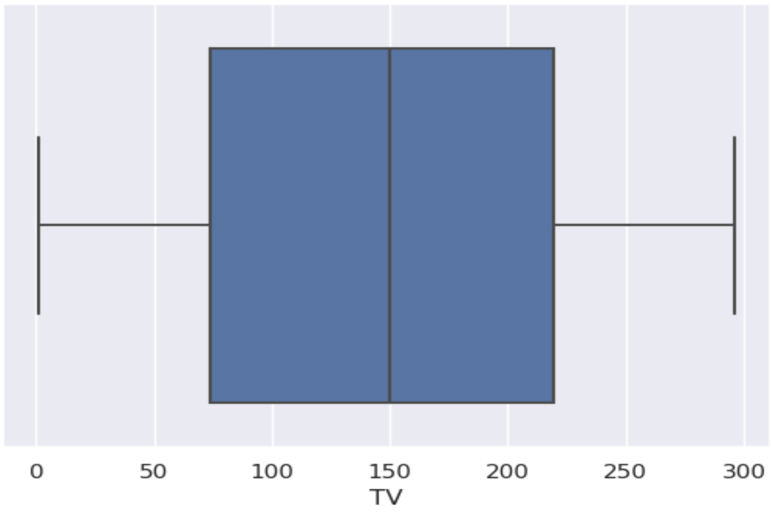


Fig 1- Histogram for TV

Based on Fig 1, the histogram appears to have a uniform distribution with no gaps or outliers. There is one peak with 19 observations between $0 and $20000, and three bins with 17 observations each, $60000-$80000, $200000-

$220000 and $220000-$240000. The data has a range of $295700. There are no clusters in the distribution.



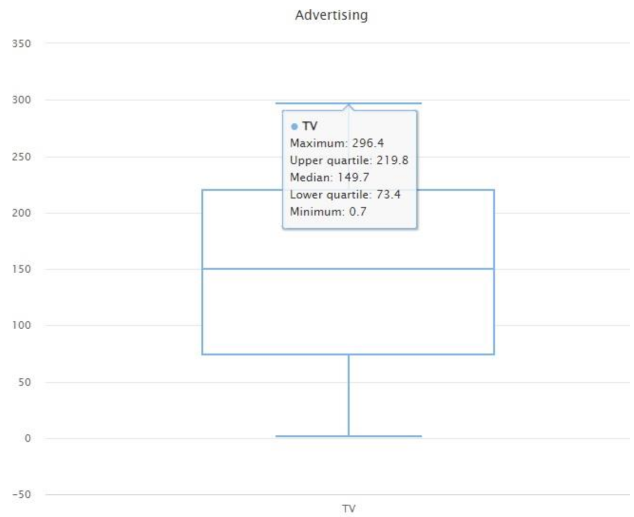


Fig 2- BoxPlot for TV

Analyzing fig 2 we can say that the distribution of the variable is slightly skewed to the right because the median $149700, is closer to the lower quartile $73400 than the upper quartile $219800, and the maximum value $296400 is far from the upper quartile. There is no outlier

Radio

Name	Type	Missing	Min	Max	Average	Deviation
radio	Real	1	-21.7	49.6	22.943	15.15

Table 2- radio

From table 2:

The variable is a continuous numeric variable represented as a real number.

There is one missing value in the radio advertising variable.

The minimum value for radio advertising is -$21700, and the maximum value is $49600.

There is only one negative value

The average or mean of radio advertising expenditure is $22943.

The standard deviation of the radio advertising variable is $15150.

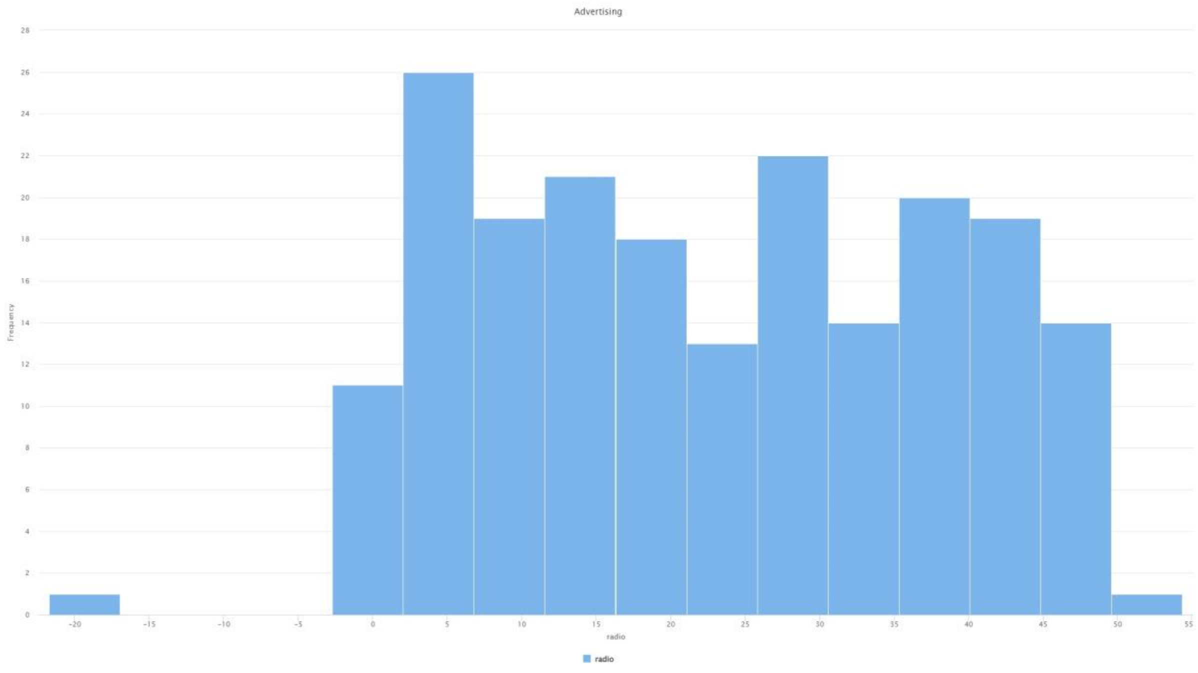


Fig 3-Histogram for radio

Looking at Fig 3, the histogram appears to have a uniform distribution with one gap of three bins between -$16947 and -$2687 and one outlier. There is one peak with 26 observations between $2067 and $6820. The data has a range of $295700. There are no clusters in the distribution.

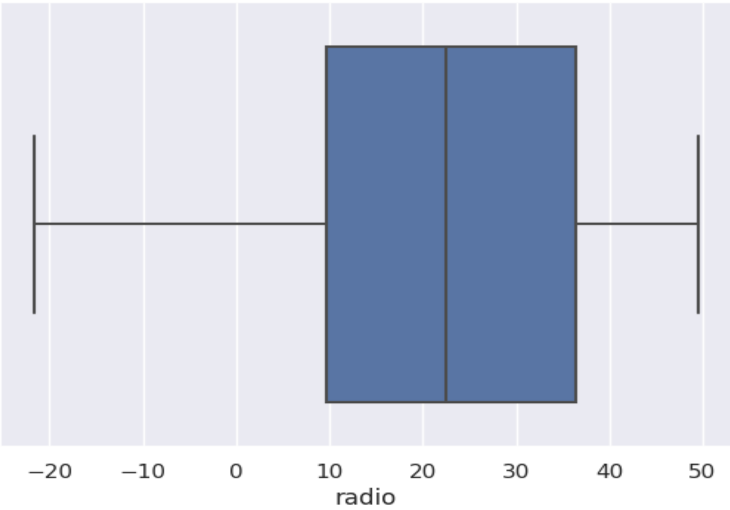


Fig 4-BoxPlot for radio

Analyzing fig 4 we can say that the distribution of the variable is skewed to the right because the median $22500, is closer to the lower quartile $9600 than the upper quartile $36500, and the maximum value is far from the upper quartile and the minimum value from the lower quartile. There is no outlier.

Newspaper

Name	Type	Missing	Min	Max	Average	Deviation
newspaper	Real	0	0.3	230.2	31.589	25.947

Table 3-newspaper

From table 3:

The variable is a continuous numeric variable represented as a real number.

There is no missing value in the newspaper advertising variable.

The minimum value for newspaper advertising is $300, and the maximum value is $230200.

There is only one negative value

The average or mean of newspaper advertising expenditure is $31589.

The standard deviation of the newspaper advertising variable is $25947.

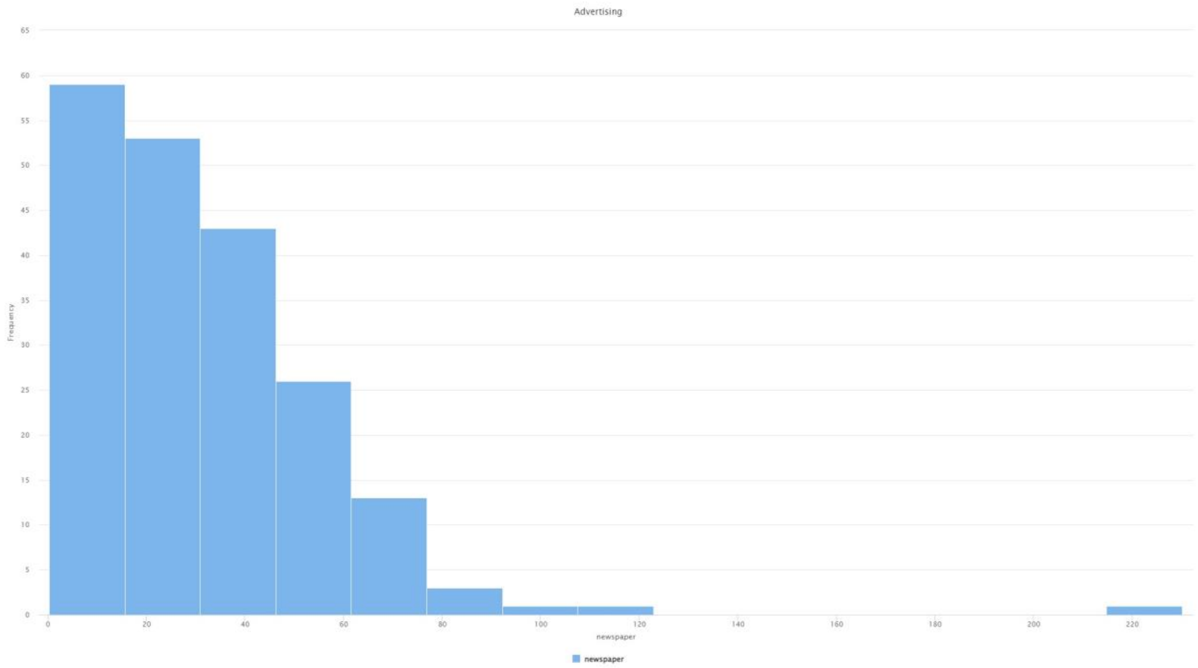


Fig 5-Histogram for newspaper

In fig 5 we can see that, the histogram have a right skewed distribution with one gap of three bins between $122913 and $214873. There is one peak with 59 observations between $300 and $15626. The data has a range of $229900.

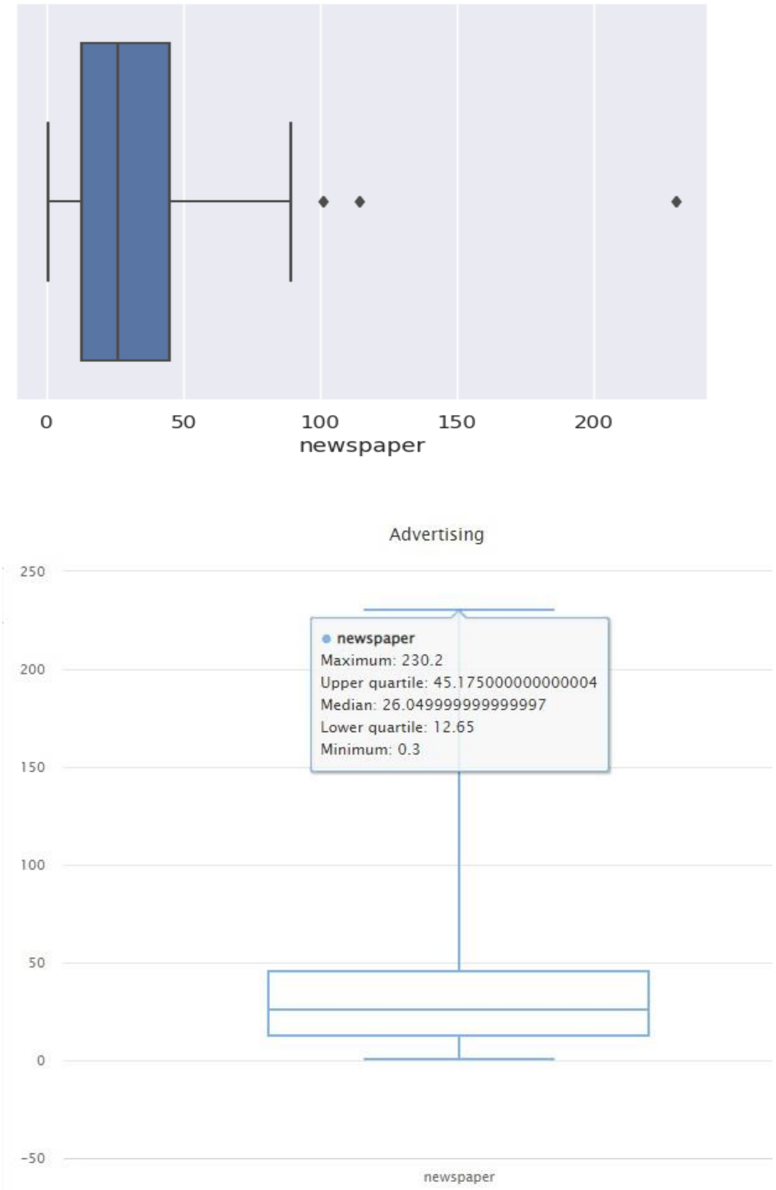


Fig 6-BoxPlot for newspaper

As can be seen in fig 6, The median is $26050, is closer to the lower quartile $12650 than the upper quartile $45175, and the maximum value is far from the upper quartile and the minimum value from the lower quartile. There are three outliers starting at approximately $100000.

Sales

Name	Type	Missing	Min	Max	Average	Deviation
sales	Real	1	1.6	27	14.065	5.196

Table 4

From table 3:

The variable is a continuous numeric variable represented as a real number.

There is one missing value in the unit sales variable.

The minimum value for sales is 1600 units, and the maximum value is 27000 units.

The average or mean of sales is 14065 units.

The standard deviation of sales variable is 5196 units .

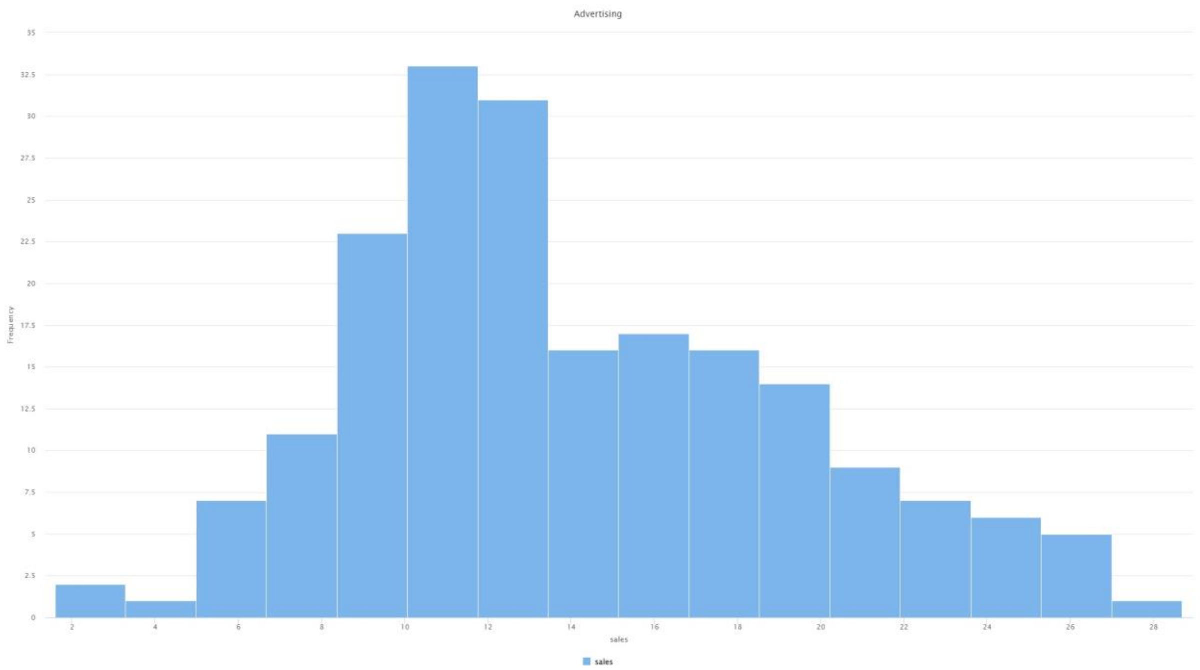


Fig 7- Histogram sales

In fig 7 we can see that, the histogram have a right skewed distribution with no gap. There is one peak with 33 observations between 10067 and 11076 units sold. The data has a range of 25400.

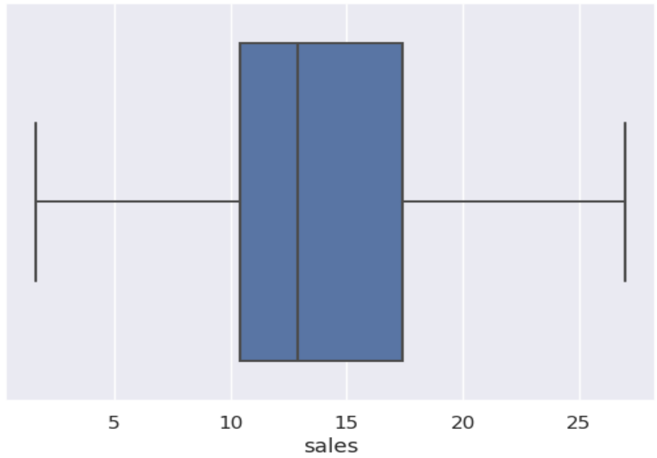


Fig 8- BoxPlot for sales

Fig 8 shows that, the median is 12900 units sold, is closer to the lower quartile 10400 units than the upper quartile 17400 units, and the maximum value is far from the upper quartile and the minimum value from the lower quartile.

Quality problems

1 -TV, radio and sales have a missing value each.

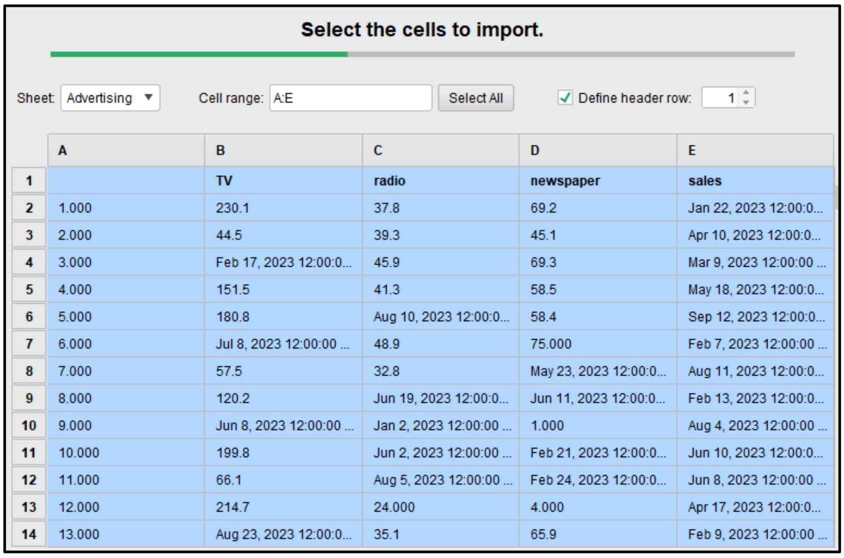
2-Radio has a negative value, but this attribute has an economic interpretation.

3-Newspaper has three outliers.

Overall, the data quality is good, with well-done datatypes, and the distributions are suitable for regression analysis. However, it is important to note that the dataset only covers a limited time period and geographical area, which may limit the generalizability of our findings.

Data preparation

Before loading the data into RapidMiner, a review was performed in an Excel sheet to correct the data types. If this correction had not been made, we would have had a data load similar to the one shown in Fig 9.



Select the cells to import.

Sheet: Advertising Cell range: A:E Select All Define header row: 1

	A	B	C	D	E
1		TV	radio	newspaper	sales
2	1.000	230.1	37.8	69.2	Jan 22, 2023 12:00:0...
3	2.000	44.5	39.3	45.1	Apr 10, 2023 12:00:0...
4	3.000	Feb 17, 2023 12:00:0...	45.9	69.3	Mar 9, 2023 12:00:00 ...
5	4.000	151.5	41.3	58.5	May 18, 2023 12:00:0...
6	5.000	180.8	Aug 10, 2023 12:00:0...	58.4	Sep 12, 2023 12:00:0...
7	6.000	Jul 8, 2023 12:00:00 ...	48.9	75.000	Feb 7, 2023 12:00:00 ...
8	7.000	57.5	32.8	May 23, 2023 12:00:0...	Aug 11, 2023 12:00:0...
9	8.000	120.2	Jun 19, 2023 12:00:0...	Jun 11, 2023 12:00:0...	Feb 13, 2023 12:00:0...
10	9.000	Jun 8, 2023 12:00:00 ...	Jan 2, 2023 12:00:00 ...	1.000	Aug 4, 2023 12:00:00 ...
11	10.000	199.8	Jun 2, 2023 12:00:00 ...	Feb 21, 2023 12:00:0...	Jun 10, 2023 12:00:0...
12	11.000	66.1	Aug 5, 2023 12:00:00 ...	Feb 24, 2023 12:00:0...	Jun 8, 2023 12:00:00 ...
13	12.000	214.7	24.000	4.000	Apr 17, 2023 12:00:0...
14	13.000	Aug 23, 2023 12:00:0...	35.1	65.9	Feb 9, 2023 12:00:00 ...

Fig 9- Inicial upload to RapidMiner

Subsequently, filters were performed on the data loaded into RapidMiner, as shown in Fig 10.

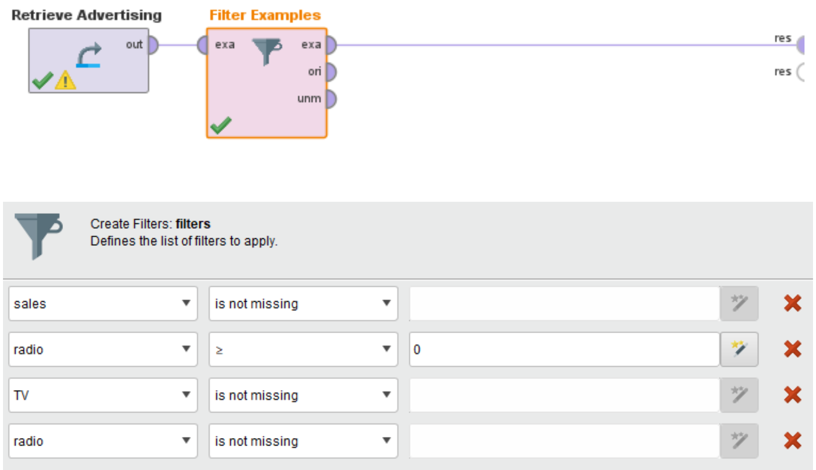


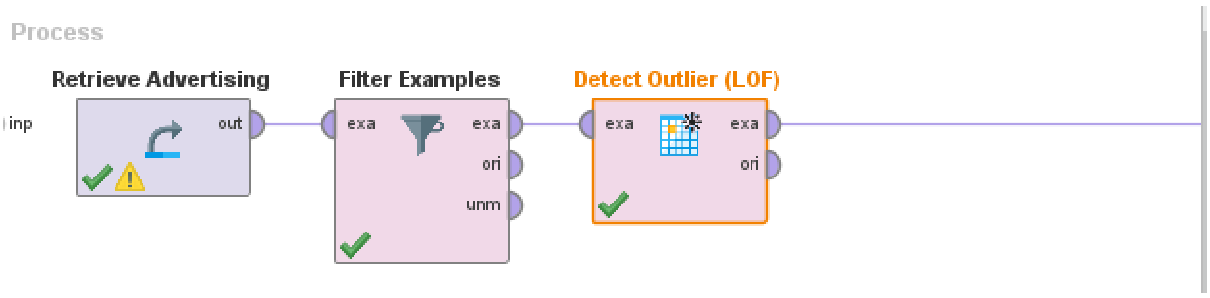
Fig 10- Filters in RapidMiner

At this stage, all observations containing missing values were discarded because the linear regression model requires complete data for accurate predictions. If the dataset contains missing values, it can lead to inaccurate parameter estimates and may affect the reliability of the model. Additionally, missing data can introduce bias in the model, especially if it is not missing at random.

While the simple linear regression proposed in the first approach section only requires discarding one row where sales is missing, and then one row in the linear regression for tv where its value is missing that could be used in the newspaper and radio regression later (the same applies the missing value in radio), discarding three rows instantly does not represent a significant percentage of observations and should still be discarded for the second approach.

Likewise, it was ensured that the values corresponding to radio advertising expenses are positive, because of the economic interpretation discussed before.

In addition, an arbitrary maximum value of 100 was assigned to newspaper advertising expenditures to rule out the outliers, after using the Detect Outlier (LOF) ("Local Outlier Factors") operator and confirming the assumptions made during data understanding, as shown in Fig 11.



Row No.	outlier ↓	id	TV	radio	newspaper	sales
106	3.995	108	90.400	0.300	230.200	8.700
17	1.826	17	67.800	36.600	114	12.500
9	1.721	9	8.600	2.100	1	4.800
100	1.614	102	296.400	36.300	100.900	23.800

Fig 11- outliers detection

As a result of these steps, a set of 193 observations was obtained and used to build the models.

Modeling

Linear regression may be a good option for this dataset because we seek to establish a linear relationship between advertising expenditure (independent variables: TV, radio and newspapers) and sales (dependent variable). Linear regression allows us to model this linear relationship and predict sales values as a function of advertising expenditure values.

In addition, it is assumed that the independent variables have a linear effect on the dependent variable and that the prediction errors are normally distributed. If these assumptions are true, the linear regression can provide a good approximation of the relationship between the variables.

First aproach

At first, it may be tempting to perform an individual linear regression for each of the advertising expenditures (TV, radio, newspapers). It is possible that this solution has been previously presented to the stackholders of the project, so we will illustrate this approach and compare it with another similar model, in order to convince the reader that, as we will see later on, the latter model produces better results.

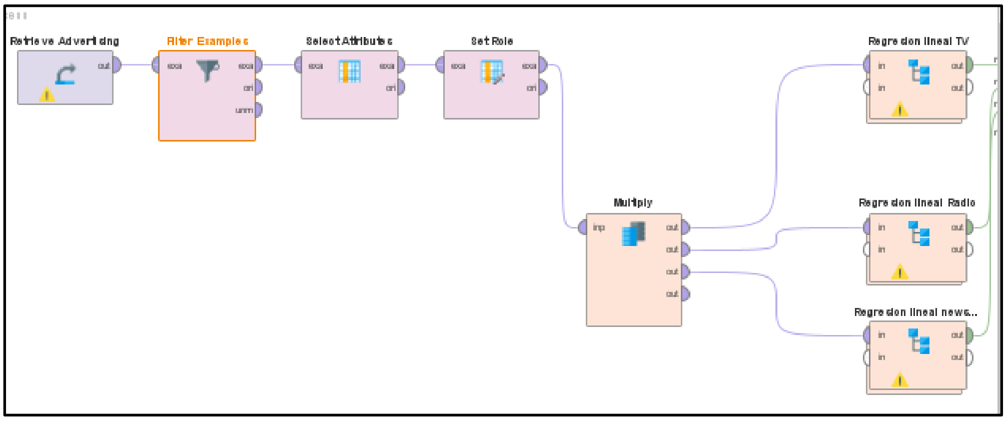


Fig 12-Diagram of RapidMiner

To begin with, the ID attribute was discarded, and the role of "label" was assigned to sales.

Then, within each sub-process we can find a diagram similar to the one presented in Fig 4.

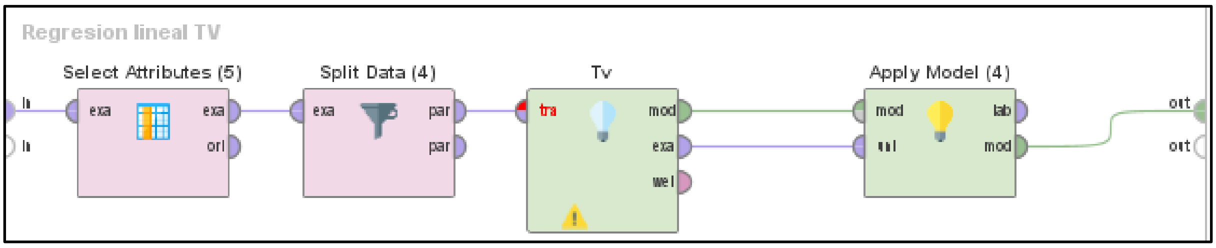


Fig 13-TV lineal regression

Where the corresponding attributes are selected (Sales and TV, in this case) and the data is divided into training (0.7) and test (0.3), although the model is never evaluated, so 100% of the observations could have been used.

The graphs corresponding to each linear regression, with its corresponding equation, are shown below:

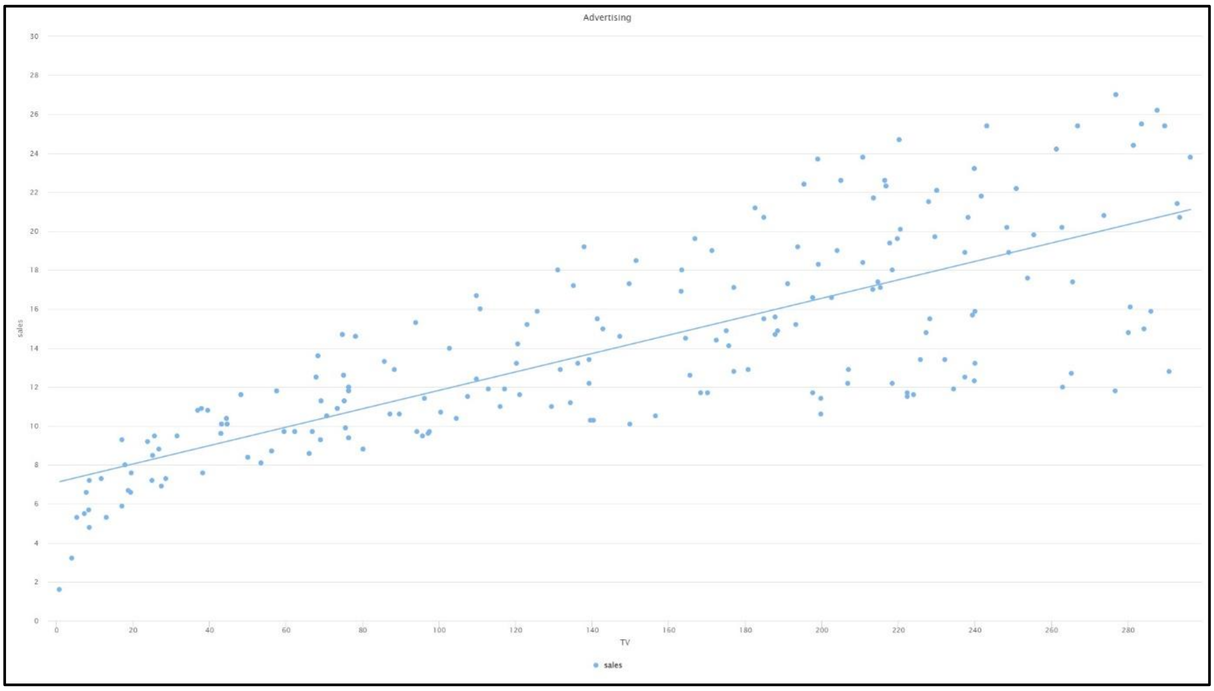


Fig 14- TV - Sales

$$\check{Y} = 0.047270919953150 16x_{TV} + 7.102184484766 35$$

The slope of the linear function for TV is 0.047, which means that for each additional unit invested in TV advertising, an increase of 0.047 units in sales is expected. Thus, an investment of $1,000 in TV advertising represents an increase of 47 units in sales. The independent term is 7.102, which indicates that when nothing is invested in TV advertising, sales are expected to be 7.102 units.

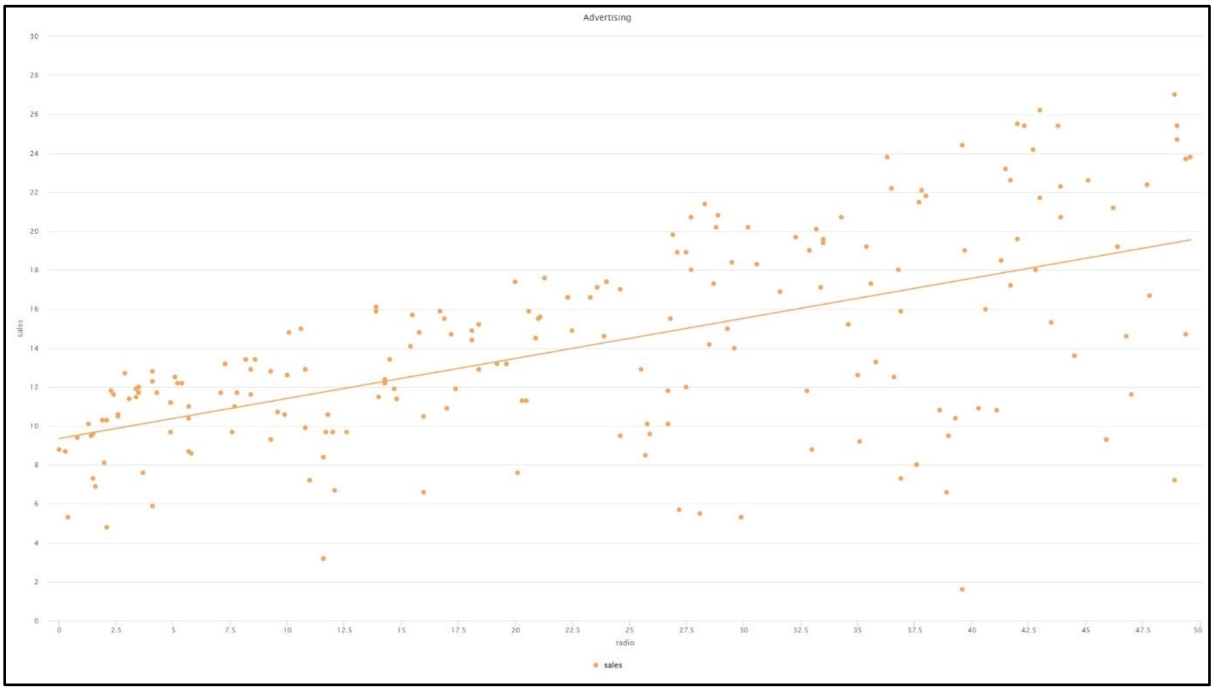


Fig 15- Radio-Sales

$$\check{Y} = 0.2055646397153556x_r + 9.357671317468274$$

The slope of the linear function for radio is 0.206, which means that for each additional unit invested in radio advertising, a 0.206 unit increase in sales is expected. Therefore, an additional investment of $1,000 represents an increase in sales of 206 units. The independent term is 9.358, indicating that when nothing is invested in radio advertising, sales are expected to be 9.358 units.

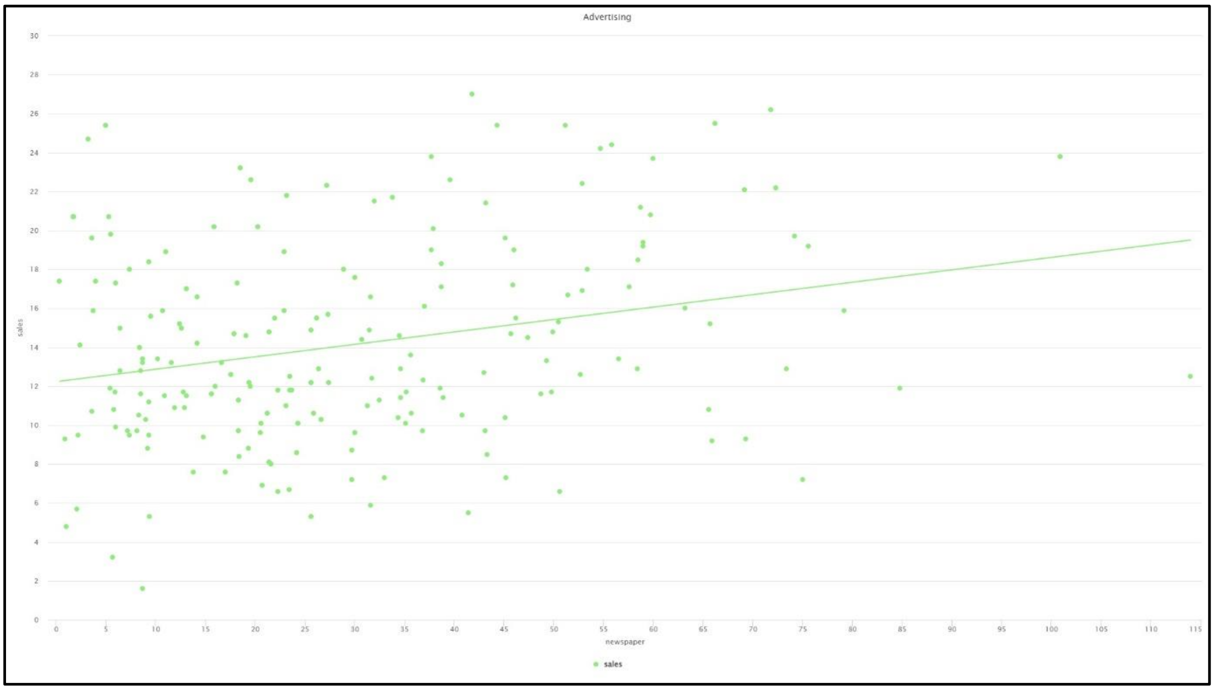


Fig 16- newspaper- Sales

$$\check{Y} = 0.06381794076750318 \mathrm{x}_{np} + 12.235148070230164$$

The slope of the linear function for the newspaper is 0.064, which means that for each additional unit invested in newspaper advertising, an increase of 0.064 units in sales is expected. That is, an increase in investment of $1,000 indicates that sales will increase by 64 units. The independent term is 12.235, indicating that when nothing is invested in newspaper advertising, sales are expected to be 12.235 units.

In this case, the slope of the function for radio is the largest, suggesting that radio advertising may have the greatest impact on sales.

Finally, let's analyze the information provided by the model for each in the following tables:

	Coefficient	Std. error	t-Stat	p-value
TV	0.047	0.0032080021285991744	14.923213624823843	0.0
intercept	7.102	0.5365955727066737	12.93206145847565	0.0

Table 5- Simple regression of sales on TV

	Coefficient	Std. error	t-Stat	p-value
Radio	0.205	0.020355189228393376	10.087124004064721	0.0
intercept	9.35	0.5627608251046562	16.62359992988953	0.0

Table 6- Simple regression of sales on radio

	Coefficient	Std. error	t-Stat	p-value
Newspaper	0.064	0.016757139542055889	3.8470759424644205	0.0001623
intercept	12.23	0.6205902690039028	19.659170178232607	0.0

Table 7- Simple regression of sales on newspaper

The standard error (std.error) is a measure of the variability of the estimated coefficient. It indicates how much the coefficient is expected to vary if the model is fitted to different data samples.

In other words, the standard error measures the precision of the estimated coefficient relative to the sample size. A smaller standard error means that the estimated coefficient is more accurate and is expected to have a smaller variation in different data samples.

In the specific case of the coefficient of the variable "TV" in this linear regression model, the value of the standard error is 0.0032. This means that the coefficient is expected to vary around 0.0032 units if the model is fitted to different data samples. That is, if new data samples are collected, the estimated coefficient of the variable "TV" could vary by about +/- 0.0032 units from the current value.

Likewise, this can be understood as the same for the standard error of radio and newspaper.

The t-statistic is a measure of how many standard deviations the estimated coefficient is from the expected value under the null hypothesis (assumption that there is no significant relationship between the predictor variables and the response variable).

In the context of a linear regression, the t-stat is calculated by dividing the value of the estimated coefficient by the standard error of the coefficient.

In general, a t-stat with an absolute value greater than 2 is considered statistically significant, suggesting that the coefficient is not zero and that the corresponding variable has a significant effect on the response variable.

In this case the t-stat value for the variable "TV" is 13.7. This value indicates that the estimated coefficient for the variable "TV" is significantly different from zero and that the variable has a significant effect on the response variable (sales). The same can be applied to the t-stat obtained in the Radio model.

The p-value is the p-value associated with the coefficient of the explanatory variable. It indicates the probability of obtaining a coefficient as extreme as the one observed if the true relationship between the variables were null (i.e., there was no relationship between TV investment and sales). In this case, the value is 0.0, suggesting that the "TV" variable is significantly relevant in predicting sales. The same can be applied to the p-value obtained in the Radio model.

For the Newspaper case the t-stat value is 3.847 indicating that the regression coefficient for "Newspaper" is significantly different from zero and the p-value is 0.0001623 in this case indicating that there is sufficient statistical evidence to reject the null hypothesis that the regression coefficient for "Newspaper" is zero.

However, it is important to keep in mind that the interpretation of these results must be done in the context of the complete dataset and variables, and must be complemented by a more detailed analysis of the quality of the model and the validity of the underlying assumptions.

Second aproach

Fitting a separate simple linear regression model for each predictor may not be the best approach when dealing with multiple predictors, as it does not take into account the potential interactions between them. In this case, it can lead to misleading estimates of the individual media effects on sales if the media budgets are correlated with each other in the data set. Therefore, using all predictors in a multiple linear regression model is a better approach as it allows for the examination of the effects of each predictor while taking into account the potential interactions between them. This would also enable a single prediction of sales given levels of all three advertising media budgets. The data used for the multiple linear regression consist of the previous 195 rows with the same treatment.

The pipeline is as the lector can see in Fig 8.

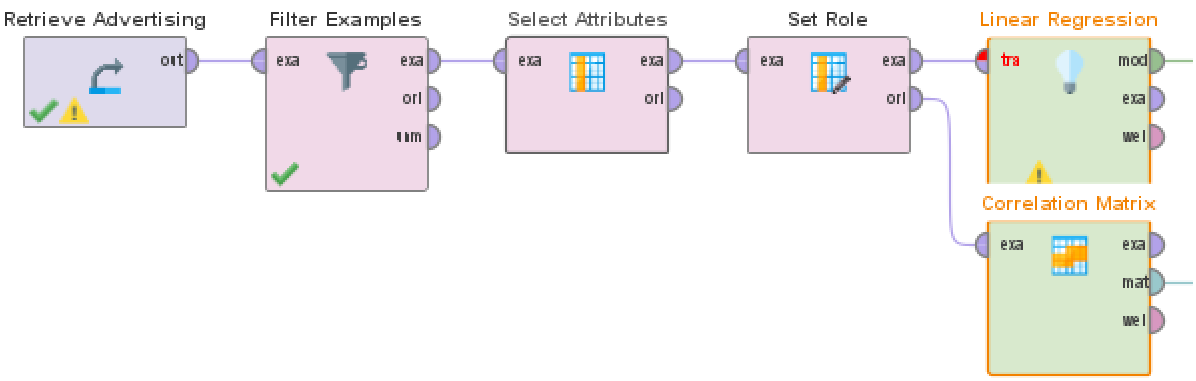


Fig 17- Multiple linear regression in RapidMiner, and a correlation matrix for the attributes

Table 4 presents the results of the multiple linear regression analysis:

	Coefficient	Std.error	t-stat	p-value
Intercept	2.93	0.3159	9.263	0.0
TV	0.045	0.0014	31.768	0.0
newspaper	0.001	0.006	0.225	0.8221

radio	0.1896	0.009	21.851	0.0

Table 8-Multiple linear regression of sales on Tv, radio, and newspaper

In Table 4, we can see the coefficient estimates obtained from a multiple regression analysis using the Advertising data to predict product sales based on TV, radio, and newspaper advertising budgets. These results indicate that increasing radio advertising spending by $1,000, while keeping TV and newspaper advertising budgets constant, is associated with an increase in sales of approximately 189.6 units. Similarly, increasing TV advertising spending by $1,000, while keeping radio and newspaper advertising budgets fixed, is associated with an increase in sales of approximately 45 units. Comparing these coefficient estimates to the ones in Tables 1 and 2, we observe that the estimates for TV and radio in the multiple regression model are similar to those obtained from simple linear regression models. However, while the coefficient estimate for newspaper in Table 3 was significantly non-zero, in the multiple regression model it is close to zero, and the associated p-value is no longer significant, with a value around 0.8221.

The fact that the coefficient estimate for newspaper in the multiple regression model is close to zero and the corresponding p-value is not significant suggests that there is no significant relationship between newspaper advertising and product sales after accounting for the effects of TV and radio advertising.

Table 5 shows the correlantion matrix for the variables. A correlation matrix is a table that shows the pairwise correlations between several variables. In other words, it displays the correlation coefficients between each variable in a dataset. Correlation coefficients measure the degree to which two variables are related.

In the context of Table 5, the correlation matrix displays the pairwise correlations between four variables: TV advertising budget, radio advertising budget, newspaper advertising budget, and product sales. Each cell in the table shows the correlation coefficient between two variables. The diagonal of the table shows the correlation of each variable with itself, which is always equal to 1.

For example, the correlation between TV advertising budget and product sales is 0.775, which indicates a strong positive relationship between these two variables. Similarly, the correlation between radio advertising budget and newspaper advertising budget is 0.347, indicating a moderate positive relationship between these two variables. By examining the correlation matrix, we can identify potential relationships between variables that may be relevant in predicting or understanding the outcome variable.

	TV	radio	newspaper	sales
TV	1.0	0.058	0.099	0.775
radio		1.0	0.347	0.587
newspaper			1.0	0.267
sales				1.0

Table 9- Correlation matrix for TV, radio, newspaper, and sales

The correlation matrix in Table 5 shows that there is a correlation between the amount spent on radio and the amount spent on newspaper advertising, with a correlation coefficient of 0.347. This means that in markets where more is spent on radio advertising, there tends to be higher spending on newspaper advertising as well. If we assume that newspaper advertising has no direct impact on sales, but radio advertising does increase sales, then in markets where we spend more on radio, our sales will tend to be higher. However, since we also tend to spend more on newspaper advertising in those same markets, a simple linear regression analysis that only examines sales versus newspaper may wrongly suggest that higher values of newspaper are associated with higher sales, when in fact, it is the effect of radio on sales that is driving this association. Therefore, newspaper sales are acting as a surrogate for radio advertising, and newspaper gets undeserved "credit" for the effect of radio on sales.

Now that it has been shown that the choice of the multiple linear regression model is better, a cross validation will be performed.

Cross-Validation for the multiple lineal regression model

RMSE (Root Mean Squared Error) will be used as metric for evaluating the performance of the multiple linear regression models. RMSE measures the average distance between the predicted values and the actual values. It is a good metric to use because it penalizes large errors more than small errors and gives an idea of how far off the predicted values are from the actual values. It provides an intuitive measure of model prediction error, considering both bias and variance. More importantly, RMSE can be compared against other models to see which one performs better in predicting the target variable. Lastly, it is a widely used metric in the field of machine learning and statistics, making it a familiar and established way to evaluate model performance.

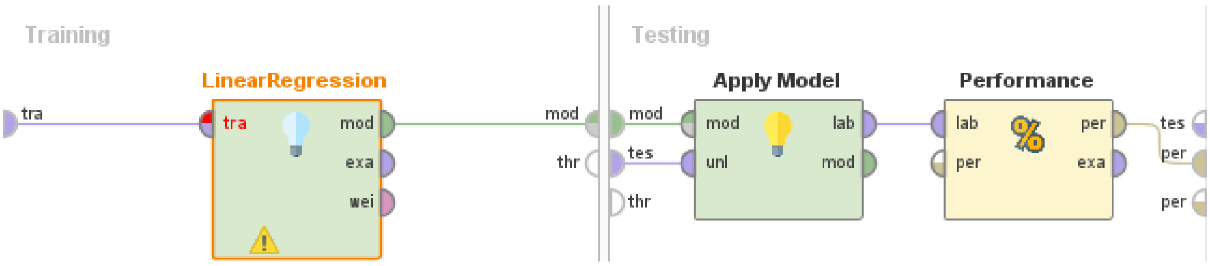


Fig 18-Cross-Validation process

Modifications were made to the parameters of the Linear regression operator which can be seen in Fig 18.

The parameters that were modified were:

-Feature selection: none, M5 prime, or greedy

-Min tolerance: 0,05 - 0,5

-Ridge: 1.0E-10 - 1.0E-1

Except for the change made between none, M5 prime/greedy, the model obtains the same results for this metric.

The parameters used for the following linear regression were:

-Feature selection:M5 prime

-Min tolerance: 0.05

-Ridge: 1.0E-8

The resulting linear regression is shown in the table below.

	Coefficient	Std.error	t-stat	p-value
Intercept	2.96	0.304	9.744	0.0
TV	0.0453	0.0014	31.449	0.0
radio	0.1903	0.0081	23.211	0.0

Table 10- Multiple linear regression of sales on Tv, radio, and newspaper after cross validation

Similar interpretations can be made for table 8. In this case the newspaper variable is discarded, obtaining a better model, which only depends on TV and radio.

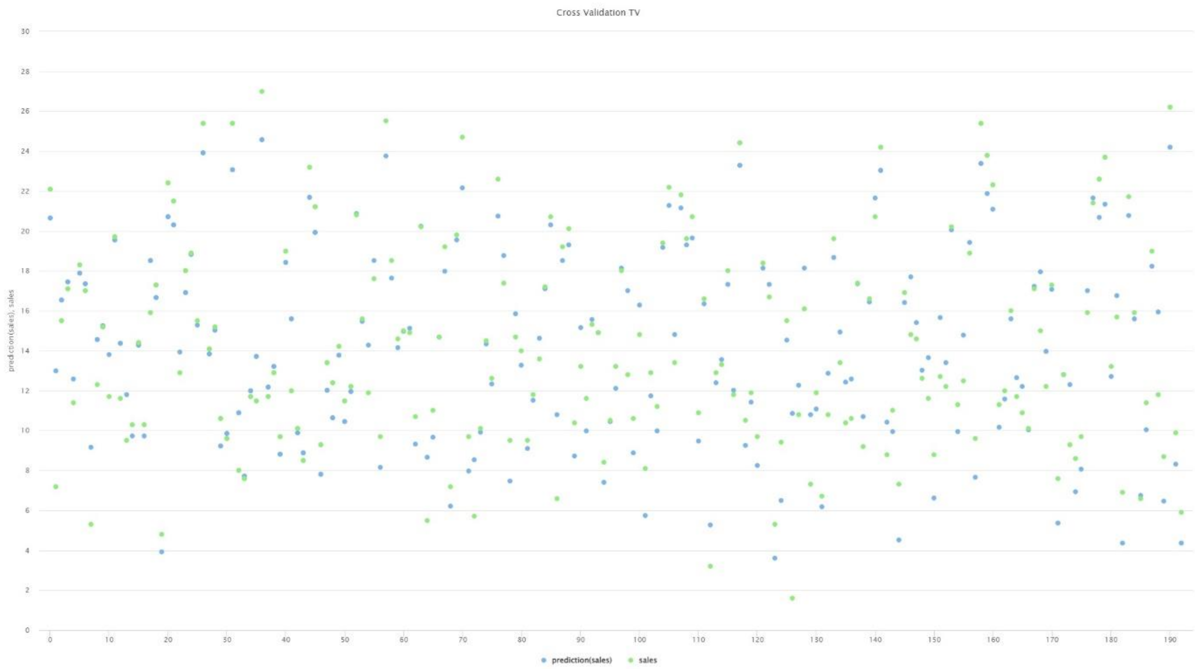


Fig 19-Prediction and sales scartter plot

Finally, in Fig. 19, we can see that the predicted values of the model are quite similar to the real values with only two variables needed.

Conclusions

There is a relationship between advertising budget and sales, as indicated by the multiple linear regression analysis. The regression analysis showed that there is a statistically significant relationship between advertising budget and sales, with TV and radio advertising budgets having a positive effect on sales, while newspaper advertising budget has little to no effect on sales. The correlation matrix also showed that there is a strong positive correlation between TV advertising budget and sales, as well as a moderate positive correlation between radio advertising budget and sales. Although the coefficient for radio advertising in the multiple linear regression model may be larger than the coefficient for TV advertising, it is important to consider the scale of the advertising budget for each medium. The coefficient represents the estimated change in sales associated with a $1,000 increase in advertising budget for that particular medium, but this does not necessarily mean that the impact of radio advertising is larger in magnitude than the impact of TV advertising.

According to the multiple linear regression analysis TV, and radio contribute to sales. The strength of their contribution varies. Spending on TV advertising has the largest positive impact on sales, followed by spending on radio advertising. Spending on newspaper advertising has a weaker positive impact on sales, but it is not statistically significant in the multiple regression model when controlling for the other two media. The latter being discarded as a variable in the final model.

Additionally, we can look at the correlation matrix, which shows the correlations between the three types of advertising. The positive correlation between TV and sales (0.78) is higher than the correlations between radio and sales (0.58) and newspaper and sales (0.23). However, we cannot conclude that there is synergy among the advertising media based solely on the correlation matrix.

Therefore, while there may be some synergy among the advertising media, more analysis is needed to determine the exact nature and magnitude of the relationships among the variables.

On the other hand, analysis of other performances could help to improve the model.

Recommendations

1-Increase investment in TV advertising as it has the largest impact on sales.

2-Maintain or slightly increase investment in radio advertising as it also has a positive impact on sales.

3-Reconsider investment in newspaper advertising, as it has a weaker positive impact on sales and is not statistically significant when controlling for the other two media.

4-Consider conducting further research to investigate potential interactions or synergies among the different advertising media.

5-Futher experimentation with different metrics of performance.

References

Gareth James, D. W. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Jiawei Jan, M. K. (2012). *Data mining concepts and techniques*. 225Wyman Street,Waltham, MA 02451, USA: Elsevier Inc.

Data

The data used in this project "advertising.csv" is in publicly accessible and no special credentials are required to access it.

https://www.kaggle.com/datasets/bumba5341/advertisingcsv