Case Study for the module of Scientific Programming and Mathematical Modelling

Student ID: 5552339

2024-05-07

Introduction

*A summary of the case

- The purpose of the study
- The outline of the study
- Methodology:
- What soYware do you use?
- Briefly describe the stages of mathema3cal modelling for this study starting from simplifying assump3on, objec3ves (can men3on independent and dependent var.) data type and cleaning, descrip3ve analysis, inference, and conclusion.

The media company Bisney aims to understand the relationship between advertising expenditures across TV, radio, and newspapers, and the sales performance of their action figure toy over the past three years. With a dataset of 200 observations, they want to uncover which advertisement channels are most influential in driving toy sales.

The purpose of this study is to analyze the effect of advertising expenditures on toy sales for Bisney's action figure product line. Specifically, the study aims to determine the most effective advertising channel among TV, radio, and newspaper. The study also aims to showcase a step-by-step mathematical modeling approach. Finally, the study addresses the strategic question of advising Bisney on how to distribute its resources among the three advertising channels to achieve a target of 30 thousand toy sales in the next period.

Outline of the study: The study begins by framing the case of Bisney, focusing on the link between advertising spending and toy sales. It progresses through preliminary data analysis, descriptive data analysis, and inferential data analysis, concluding with a summary of findings and recommendations.

Methodology: Using RStudio, the study simplifies assumptions and identifies independent variables (TV, radio, newspaper spending) and the dependent variable (toy sales). Data cleaning ensures dataset integrity, followed by descriptive analysis for understanding variable distributions. Inferential analysis uses correlation and regression to show relationships, guiding strategic decisions for Bisney's marketing efforts.

Main Body

Definitions

- definitions of terms
- central limit theorem, measure of central tendency, measure of spread, population and samples, linear regression

Preliminary data analysis

###type&clean * Summarize the variables and data types. * Clean the data, check for missing values.

```
## Rows: 200 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (8): TV, Radio, Newspaper, Sales, TR, TN, RN, TRN
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
## spc_tbl_ [200 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ TV       : num [1:200] 388.9 40.6 312 113.8 194.8 ...
##  $ Radio    : num [1:200] 45.4 10.2 33 38.9 10.8 ...
##  $ Newspaper: num [1:200] 33.47 24.898 3.309 0.494 51.555 ...
##  $ Sales    : num [1:200] 14.9 12.6 12.2 18.1 11.8 ...
##  $ TR       : num [1:200] 434.4 50.7 345 152.7 205.6 ...
##  $ TN       : num [1:200] 422.4 65.4 315.3 114.3 246.3 ...
##  $ RN       : num [1:200] 78.9 35.1 36.3 39.4 62.4 ...
##  $ TRN      : num [1:200] 467.8 75.6 348.3 153.2 257.2 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   TV = col_double(),
##   ..   Radio = col_double(),
##   ..   Newspaper = col_double(),
##   ..   Sales = col_double(),
##   ..   TR = col_double(),
##   ..   TN = col_double(),
##   ..   RN = col_double(),
##   ..   TRN = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

The dataset comprises four variables: TV, Radio, Newspaper, and Sales. Each variable is represented as numeric data type.The dataset has been reviewed for missing values, and there are no missing values present across any of the variables, so no data cleaning procedures are required.

###clt

To check the suitability of data using Central Limit Theorem, comparing histograms with normal distribution curve can show whether the data is normally distributed and suitable.

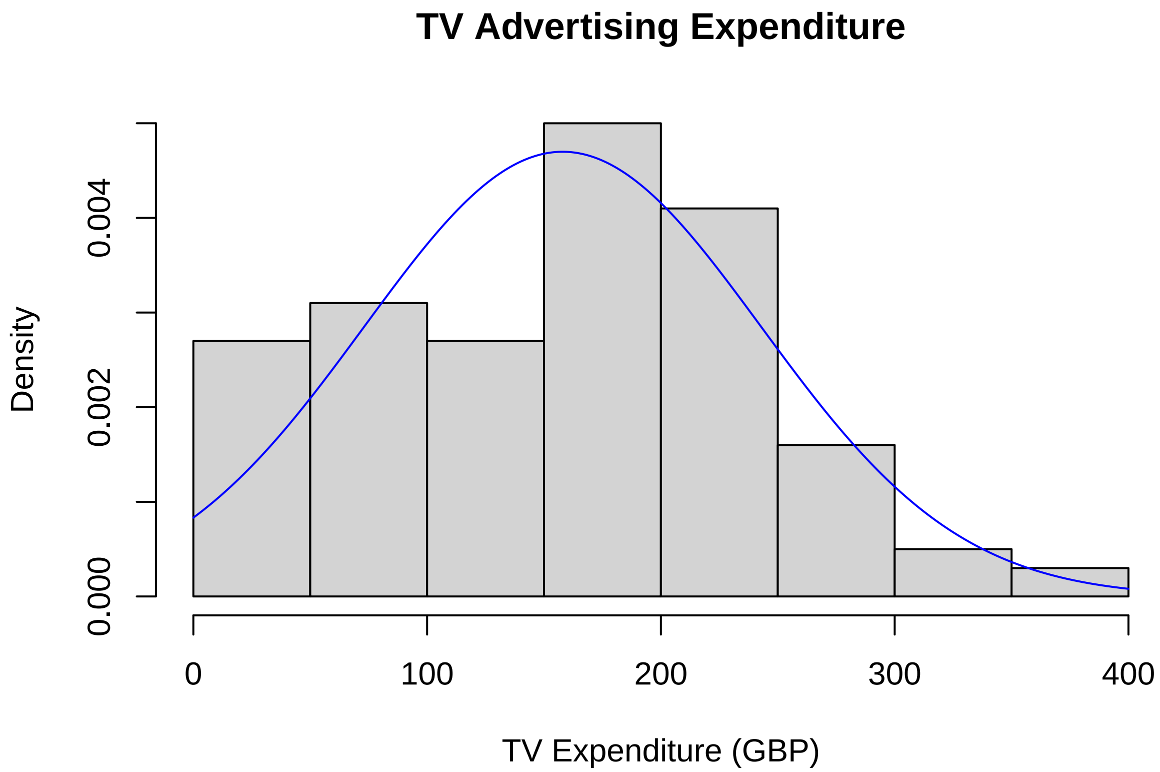


Figure1 shows that the distribution of TV advertising expenditure data is left-skewed, but is still approximately normal. Thus this group of sample data can be used in the study.

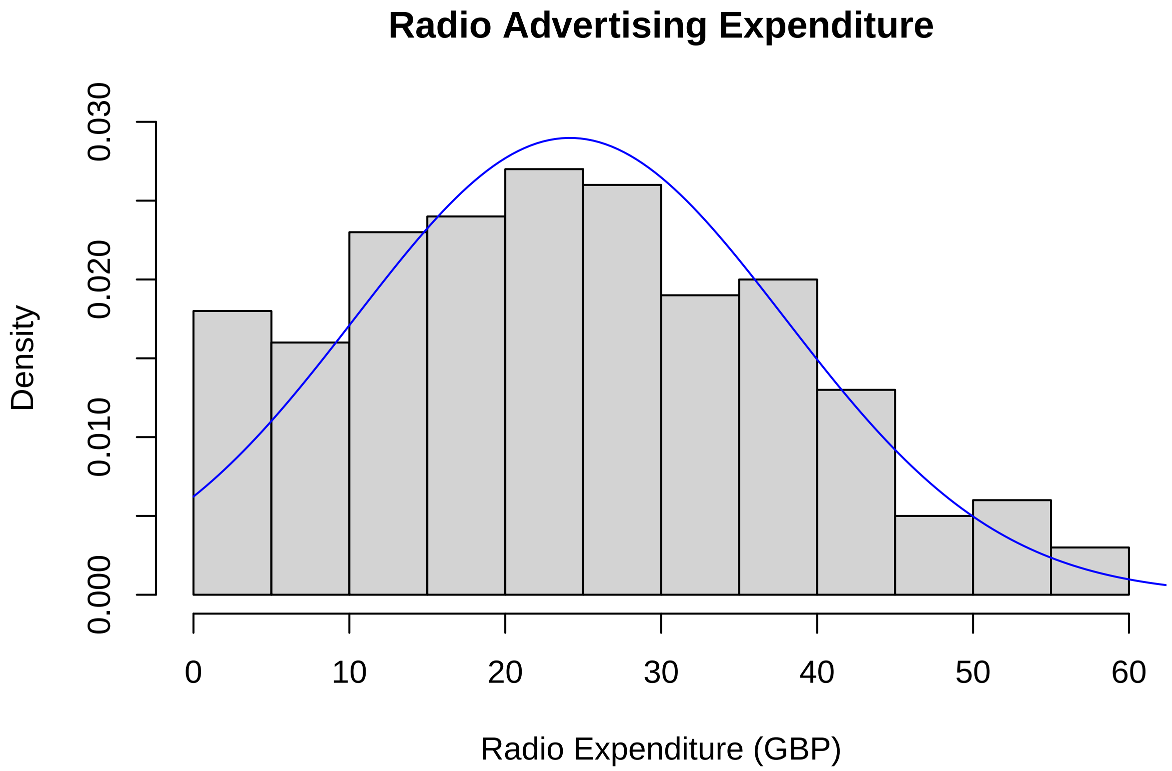


Figure2 shows that the distribution of radio advertising expenditure data is left-skewed, but is still approximately normal. Thus this group of sample data can be used in the study.

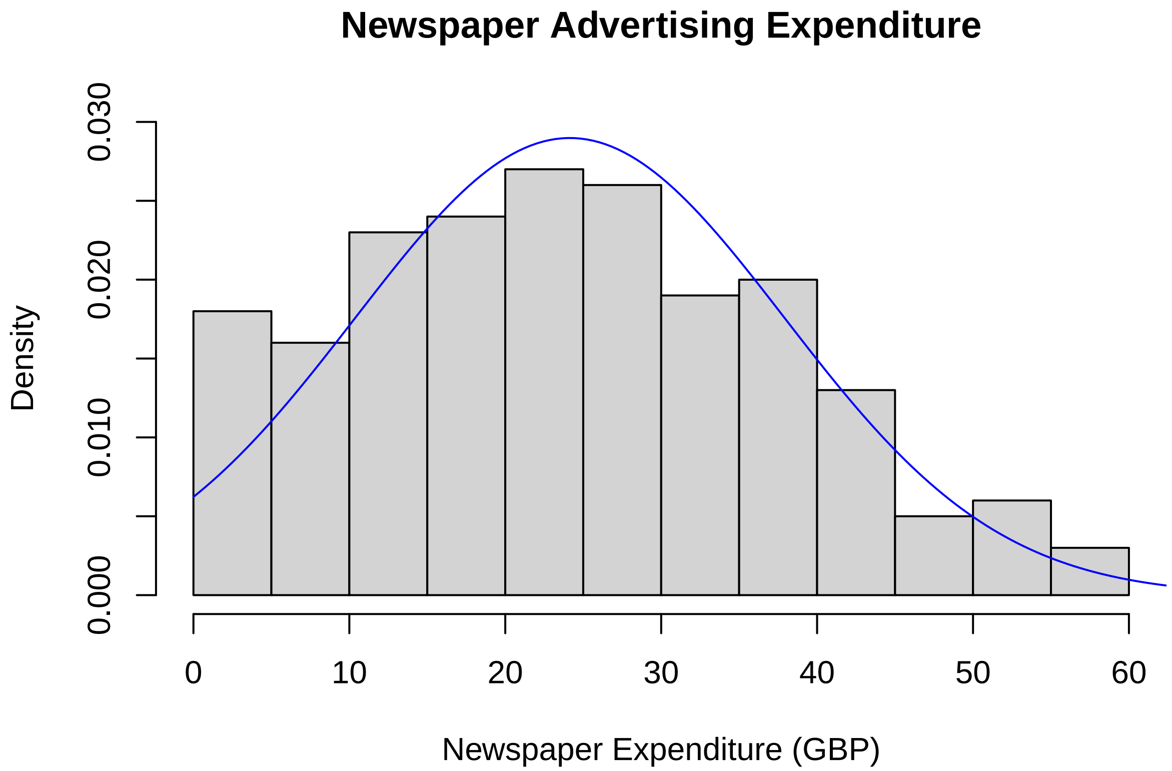


Figure3 shows that the distribution of newspaper advertising expenditure data is left-skewed, but is still approximately normal. Thus this group of sample data can be used in the study.

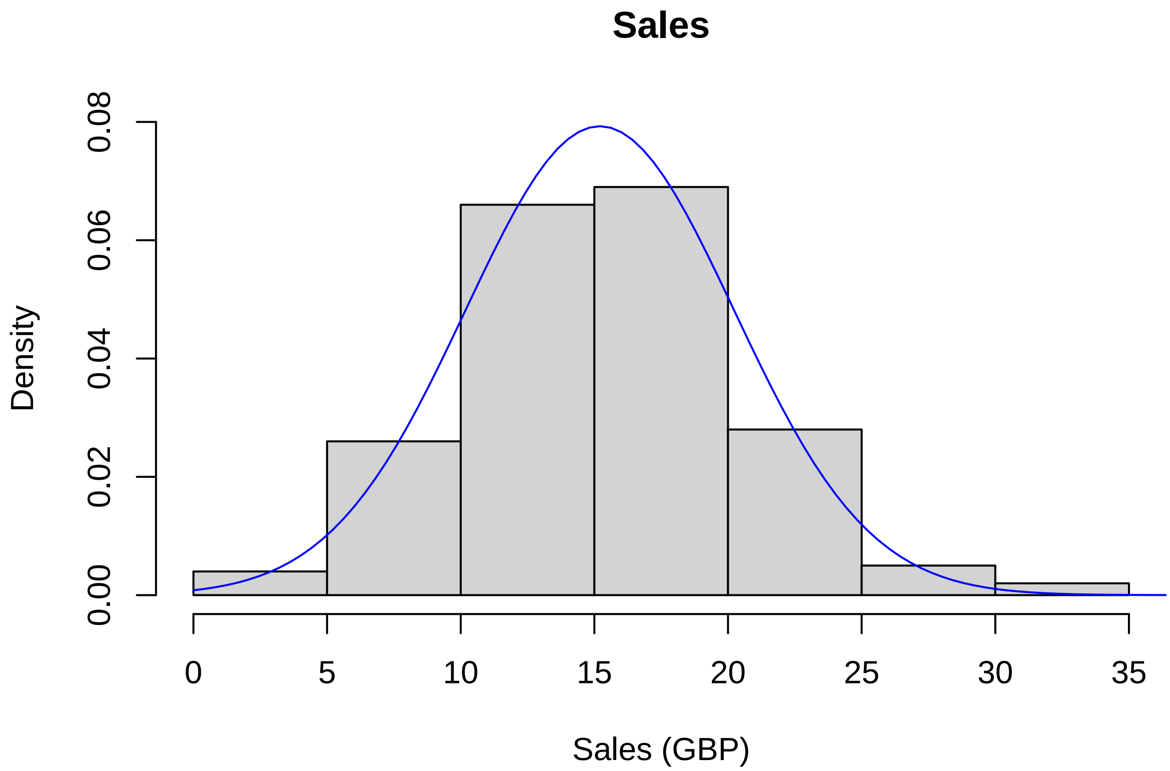


Figure4 shows that the distribution of radio advertising expenditure data is symmetrical. Therefore this group of sample data follows normal distribution and can be used in the study.

Descriptive data analysis

1. **TV Advertising Expenditure**:

```
## [1] 165.6937
```

```
## [1] 388.2614
```

From Figure1, TV advertising expenditure has a left-skewed distribution. Thus, median is suitable to measure the central tendency since it won't be affected by outliers. The median is 165.6937 GBP, showing the middle value of the data. Since standard deviation can also be affected by outliers, range is suitable for the measure of spread. The range of this group of data is 388.2614 GBP, indicating the observed TV advertising expenditure values span a large range of 388.2614 GBP.

2. **Radio Advertising Expenditure**:

```
## [1] 23.82003
```

```
## [1] 58.35869
```

From Figure 2, Radio advertising expenditure also has a left-skewed distribution. Thus the central tendency can be measured by median, which is 23.82003 GBP. Similarly, range is suitable for the measure of spread, which is 58.35869 GBP and is the span of observed Radio advertising expenditure.

3. **Newspaper Advertising Expenditure**:

```
## [1] 30.48269
```

```
## [1] 85.39051
```

From Figure 3, Newspaper advertising expenditure also has a left-skewed distribution. Thus the central tendency can be measured by median, which is 30.48269 GBP. Similarly, range is suitable for the measure of spread, which is 85.39051 GBP and is the span of observed Radio advertising expenditure.

4. **Sales**:

```
## [1] 15.20364
```

```
## [1] 5.032199
```

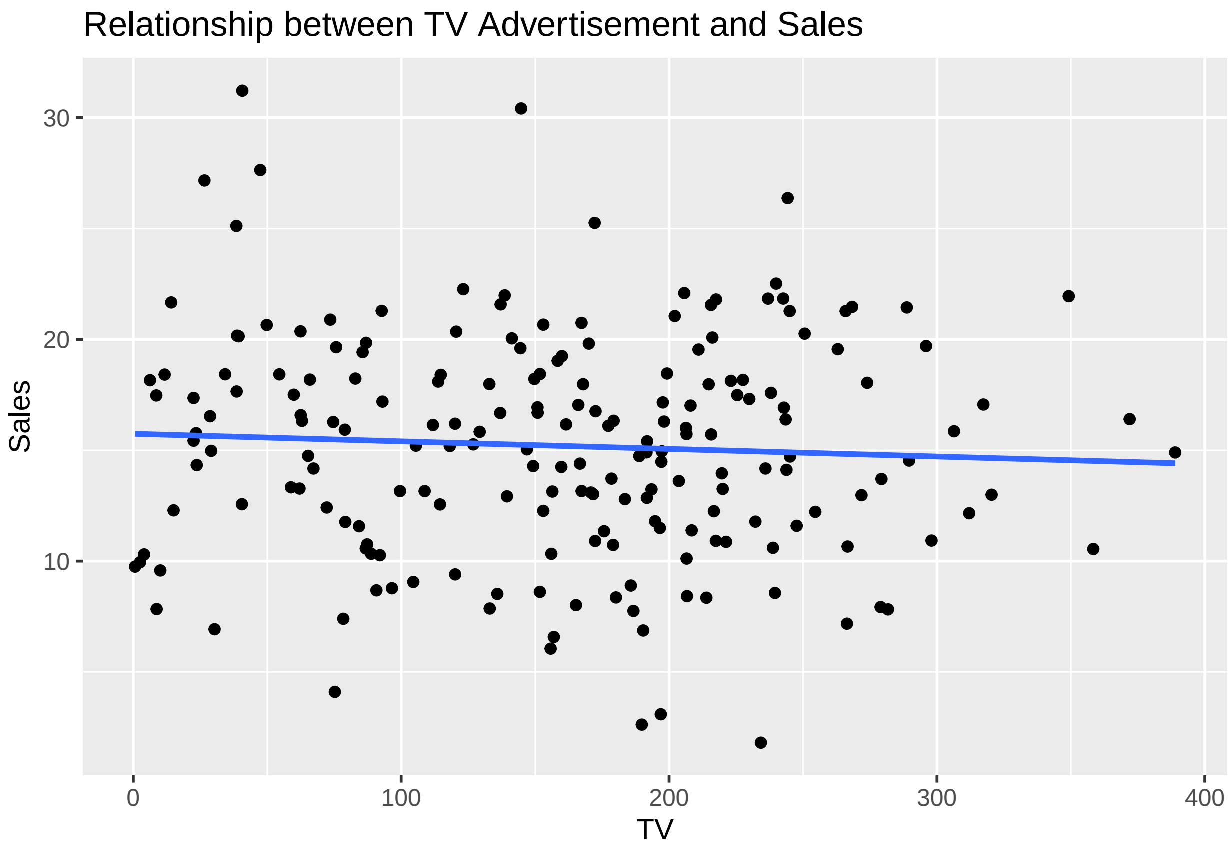
From Figure 4, the toy sales follows a normal distribution. The mean is appropriate for measure of central tendency because the distribution is relatively symmetrical and it provides a balanced representation of the data. The mean is 15.20364 GDP, showing the average level of toy sales in this period. Likewise, standard deviation is suit for the measure of spread since it can show the variability more accurately then range and there are few outliers. The standard deviation indicates that on average, toy sales in the dataset deviate from the mean (average sales) by approximately 5.032199 GDP.

Inferential data analysis

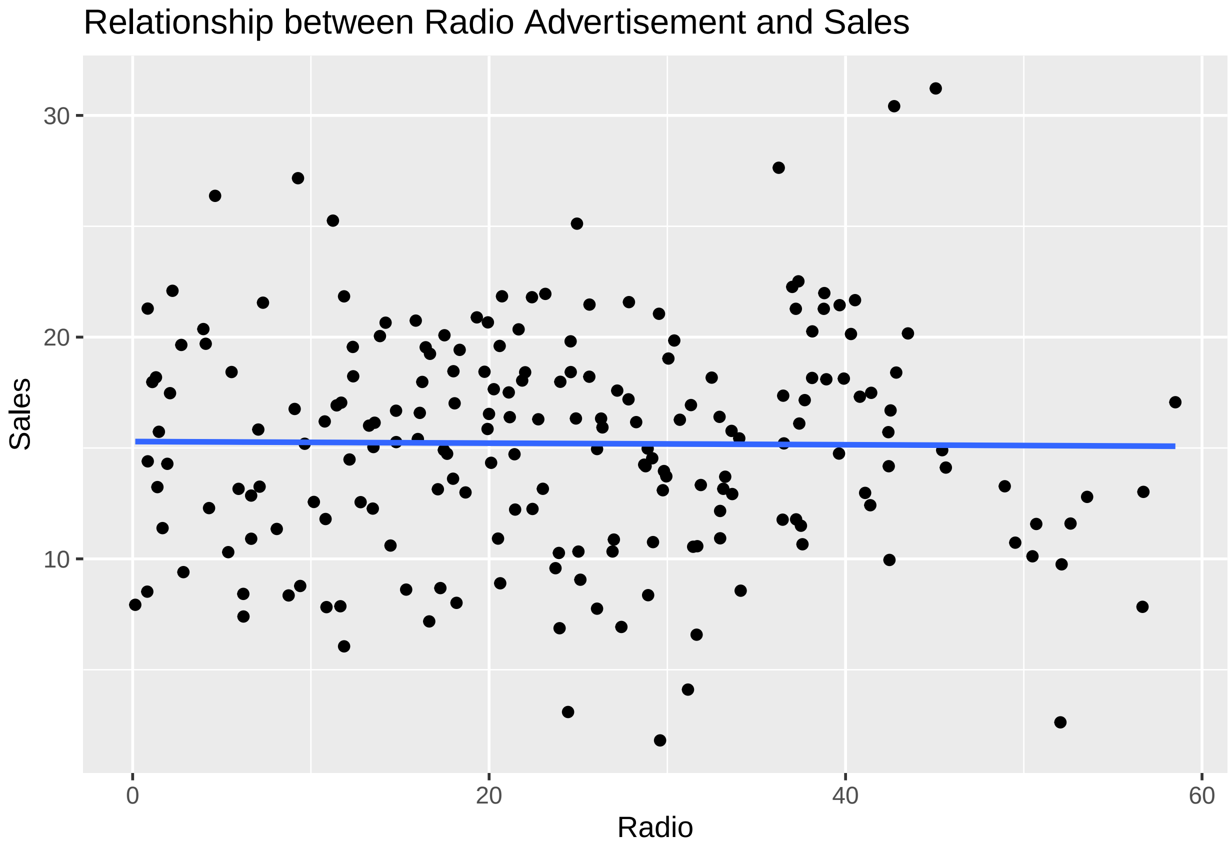
- Perform the correlation analysis between TV-Sales, Radio-sales, and newspaper-sales and analyse which of the adver3sement has the most impact to the number of sales

By performing the correlation analysis between TV-Sales, Radio-sales, and newspaper-sales, scatter graphs are drawn as follows:

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
##
## Call:
## lm(formula = Sales ~ TV, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1320  -3.4463   0.0317   3.0757  15.6135
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.743459   0.753723  20.888   <2e-16 ***
## TV          -0.003417   0.004205  -0.813    0.417
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.036 on 198 degrees of freedom
## Multiple R-squared:  0.003324,   Adjusted R-squared:  -0.00171
## F-statistic: 0.6604 on 1 and 198 DF,  p-value: 0.4174
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



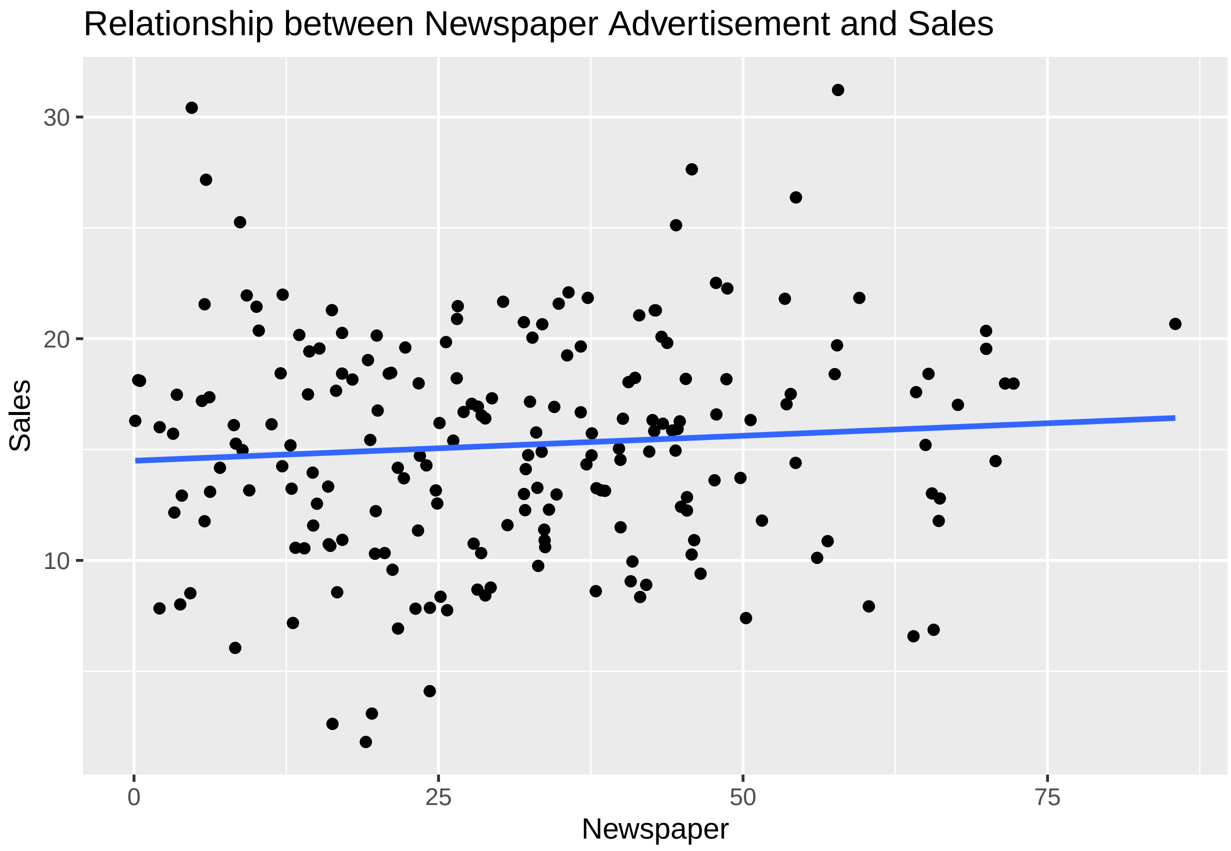
```
## [1] 25.19396
```

```
## [1] 5.019358
```

```
##
## Call:
## lm(formula = Sales ~ Radio, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3732  -3.4684   0.1092   3.1672  16.0897
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.290720   0.721560  21.191   <2e-16 ***
## Radio       -0.003607   0.025978  -0.139     0.89
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 5.045 on 198 degrees of freedom
## Multiple R-squared:  9.734e-05,  Adjusted R-squared:  -0.004953
## F-statistic: 0.01928 on 1 and 198 DF,  p-value: 0.8897
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



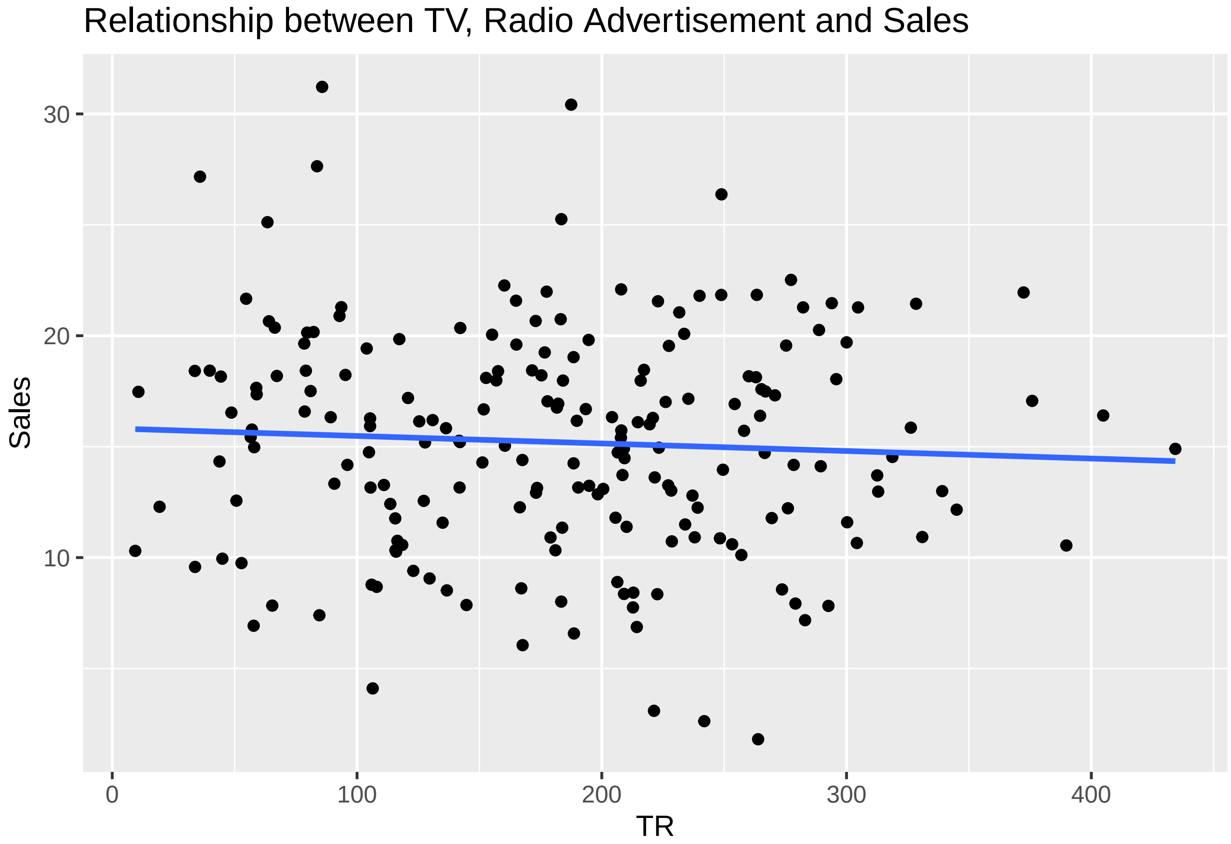
```
##
## Call:
## lm(formula = Sales ~ Newspaper, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1132  -3.6137   0.3698   3.4602  15.8153
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.49448    0.71122  20.380   <2e-16 ***
## Newspaper    0.02256    0.01960   1.151    0.251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.028 on 198 degrees of freedom
## Multiple R-squared:  0.00665,    Adjusted R-squared:  0.001633
## F-statistic: 1.325 on 1 and 198 DF,  p-value: 0.251
```

The regression functions are $y = 15.74 - 0.0034x$, $y = 15.29 - 0.0036x$, $y = 14.49 + 0.0226x$ respectively. Combined the graphs, the gradients of the regression lines show that only newspaper advertising expenditure has positive correlation with the toy sales and has the biggest impact on it.

- From your finding above, decide which of the variables you would include in your linear regression.

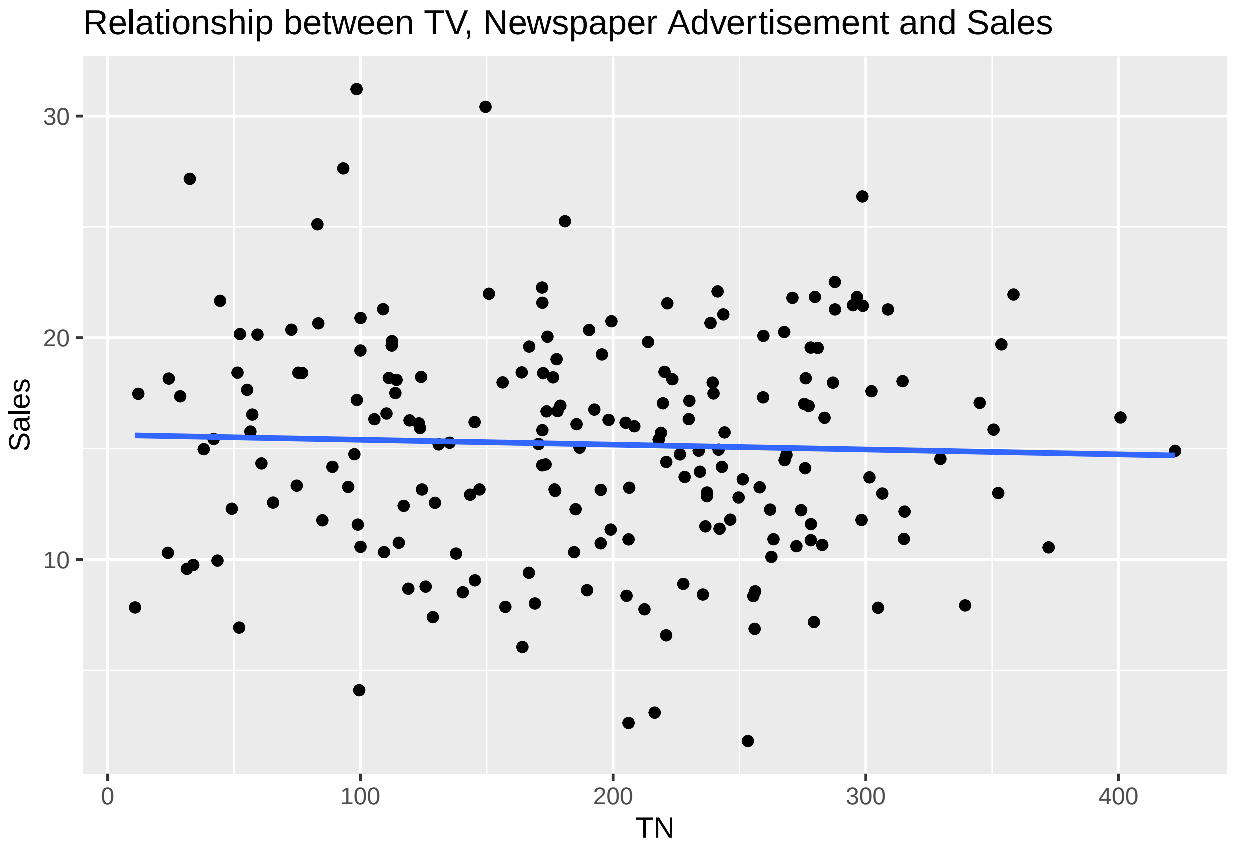
In the linear regression models, except for the four variables referred in the previous part, the independent variables also need contain the sum of TV and radio advertising expenditure, the sum of TV and newspaper's expenditure, the sum of radio and neswpaper's expenditure and the sum of all the advertising expenditure. These four regression models are as following:

```
## `geom_smooth()` using formula = 'y ~ x'
```



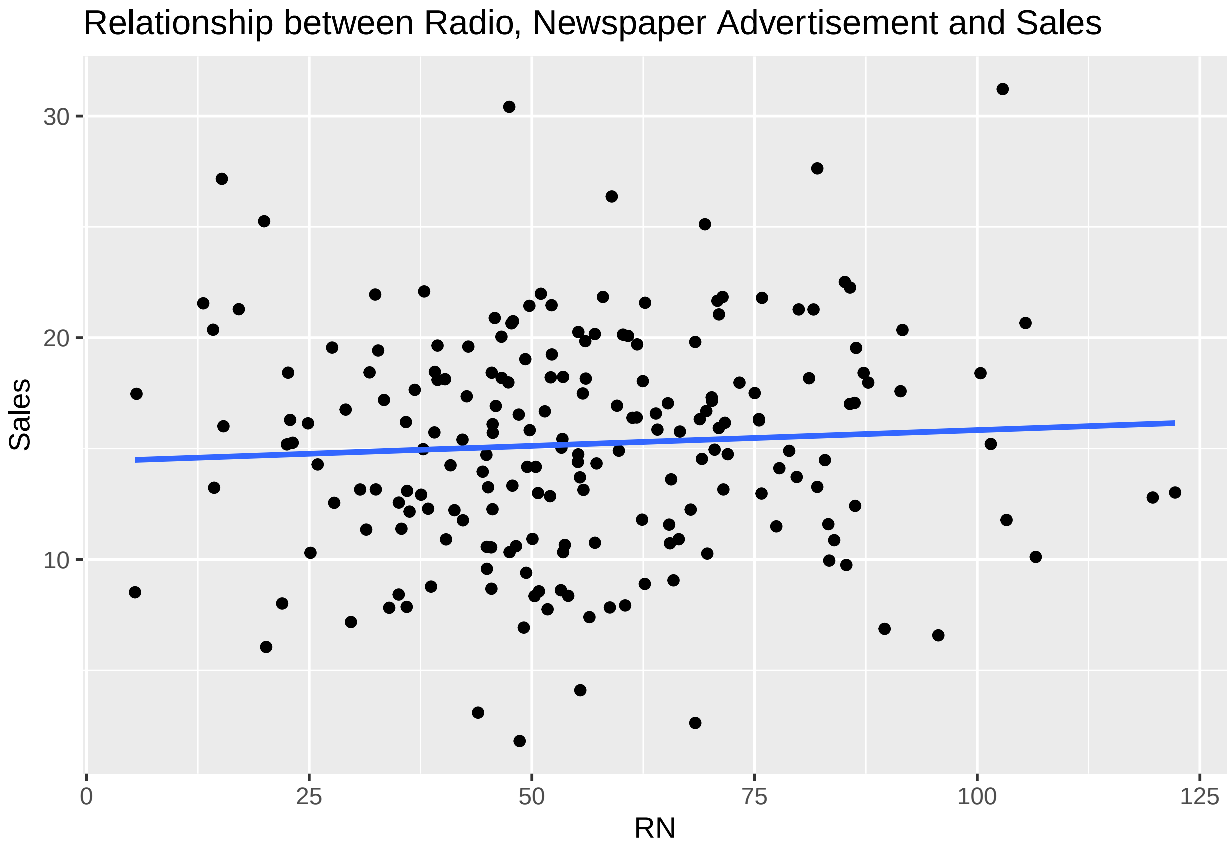
```
##
## Call:
## lm(formula = Sales ~ Radio, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3732  -3.4684   0.1092   3.1672  16.0897
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.290720   0.721560  21.191   <2e-16 ***
## Radio       -0.003607   0.025978  -0.139     0.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.045 on 198 degrees of freedom
## Multiple R-squared:  9.734e-05,  Adjusted R-squared:  -0.004953
## F-statistic: 0.01928 on 1 and 198 DF,  p-value: 0.8897

## `geom_smooth()` using formula = 'y ~ x'
```



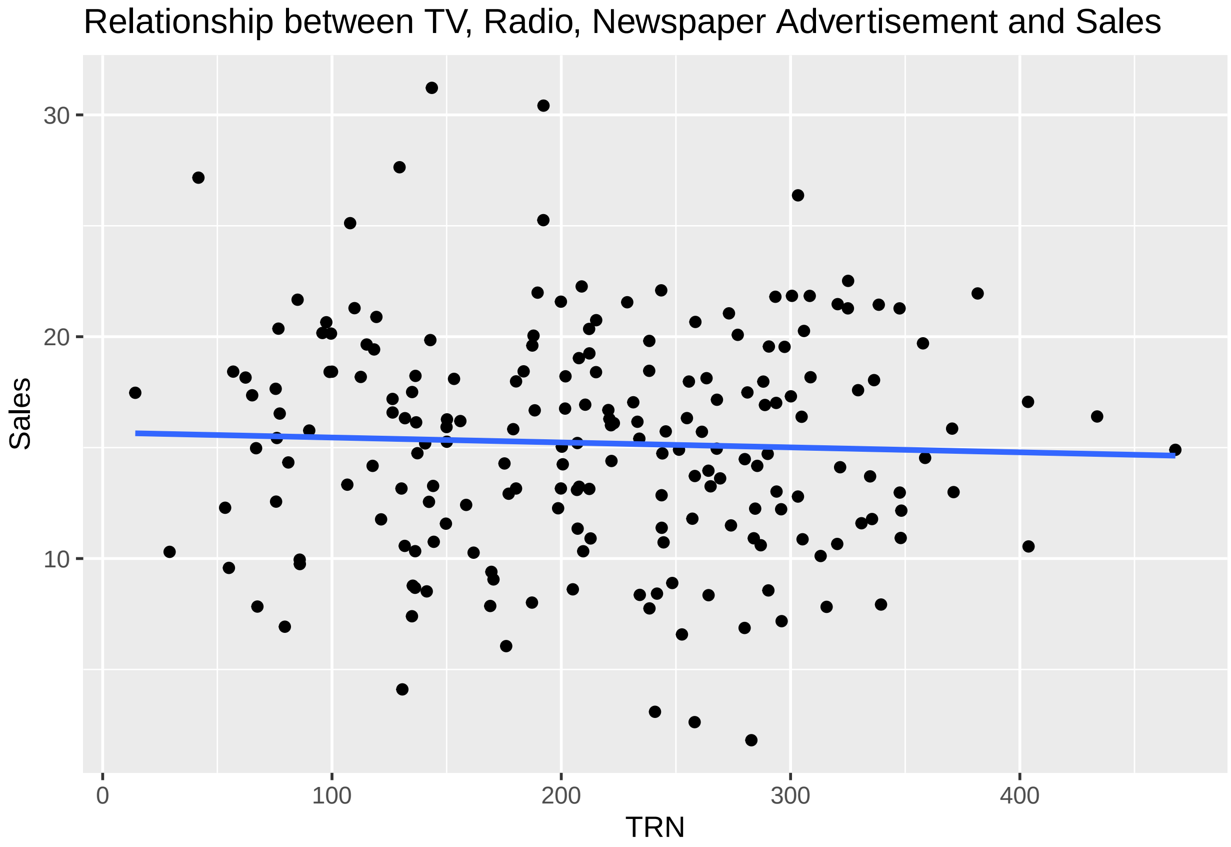
```
##
## Call:
## lm(formula = Sales ~ TN, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2525  -3.4640   0.0854   3.0295  15.8148
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.619396   0.844213  18.502   <2e-16 ***
## TN          -0.002195   0.004041  -0.543    0.588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.041 on 198 degrees of freedom
## Multiple R-squared:  0.001488,   Adjusted R-squared:  -0.003554
## F-statistic: 0.2952 on 1 and 198 DF,  p-value: 0.5875
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
##
## Call:
## lm(formula = Sales ~ RN, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2941  -3.5163   0.3999   3.1304  15.3421
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.41374    0.97559   14.77   <2e-16 ***
## RN           0.01421    0.01634    0.87    0.386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.035 on 198 degrees of freedom
## Multiple R-squared:  0.003805,   Adjusted R-squared:  -0.001226
## F-statistic: 0.7563 on 1 and 198 DF,  p-value: 0.3855
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
##
## Call:
## lm(formula = Sales ~ TRN, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2387  -3.4070   0.1171   3.0560  15.8588
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.678156   0.922907  16.988   <2e-16 ***
## TRN         -0.002222   0.003987  -0.557    0.578
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.041 on 198 degrees of freedom
## Multiple R-squared:  0.001567,   Adjusted R-squared:  -0.003476
## F-statistic: 0.3107 on 1 and 198 DF,  p-value: 0.5779
```

- Assess your regression model- their significance and accuracy.

The p-values of the regression model is 0.251, which is larger then 0.05. This means the three advertising may not have significant impacts on sales. The multiple R-squared is 0.00665 and the adjusted R-squared is 0.001633, while both show the regression model is not accurate enough.

- Answer the ques3on as to how the company could poten3ally achieve the 30 thousand sales of toy

To achieve the 30 thousand sales of toy, the company can focus on the newspaper advertising. By using the third regression function $y = 14.49 + 0.0226x$, the company would need to allocate approximately £1,327,687.61 for Newspaper advertising.

Conclusion

- Summarize the problems and the findings.
- Give some possible limita3ons of your analysis and how future analysis or decision maker can take your recommenda3on with a more informed approach