BRSM Report 1

Descriptive Statistics and Visualization of Advertising Dataset

21 February 2025

Name Vinit Mehta
Roll Number 2022111001
Mini Project 5 (Advertising Dataset)

1 Introduction

Name of Dataset: Advertising Dataset
Link to Dataset: kaggle
About Dataset:

- This dataset shows the revenue generated from sales based on the amount spent on different types of advertisements.
- The dataset contains 3 features *TV, Radio, Newspaper* and target variable *Sales*.
- Features *TV, Radio, Newspaper* are in unit of 1000$ and target variable *Sales* in M$.
- Total number of datapoints = 200 and there are no missing data in any of the columns.
- **Assumptions**:
  1. The data is for different companies around the same time period.
  2. All the companies belong to same sector/industry.

Background: Advertising has played a pivotal role in commerce since ancient times, evolving into a dominant force in capitalist economies by the mid-19th century, largely driven by newspapers and magazines [1]. The 20th century witnessed a rapid expansion of advertising across emerging media, including direct mail, radio, television, the internet, and mobile platforms. In the United States, advertising expenditure consistently averaged 2.2% of the Gross Domestic Product (GDP) between 1919 and 2007 [1].

The progression of advertising mediums follows a historical trajectory. The first newspaper advertisement in the United States appeared in the *Boston News-Letter* in 1704 [2], the first radio advertisement aired on August 22, 1922 [3], and the first paid television advertisement was broadcast on July 1, 1941, over *New York's WNBT station* [4]. The 1990s ushered in the era of digital marketing, leveraging search engines and database-driven customer targeting [5].

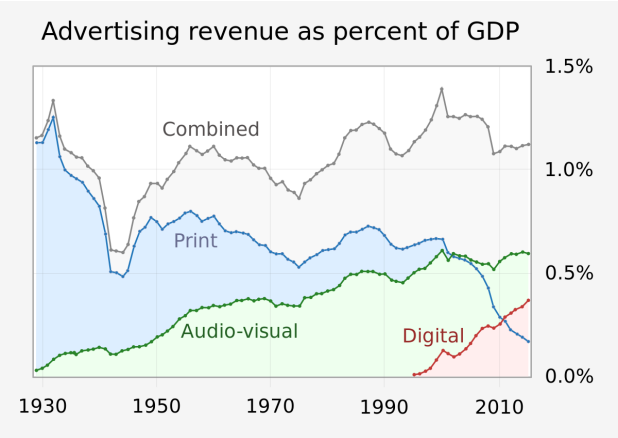


Figure 1: Advertising revenue as a percentage of US GDP (1929–2010) [1].

This dataset comprises 200 observations and explores the relationship between advertising expenditures across various media and sales performance. Historically, as new advertising mediums gained prominence, businesses adapted their spending patterns to optimize outreach and revenue generation. The dataset, created around 2018–2019, offers valuable insights into advertising expenditure trends and their impact on sales during that period.

Advertising is instrumental in driving product awareness and influencing consumer behavior. Businesses allocate budgets across multiple advertising channels such as television (TV), radio, and newspapers to maximize their market reach and sales potential. Understanding the impact of these expenditures on sales is critical for refining marketing strategies and enhancing return on investment (ROI). This study analyzes the dataset using various visualization techniques, descriptive statistics, and inferential statistical methods to determine the most effective advertising channel and provide strategic insights into advertising optimization.

2 Methods

In this section, we outline the techniques used for analyzing the dataset, including their mathematical foundations, applications, and relevance to our study.

2.1 Pie Chart

A pie chart visually represents proportions within a dataset by displaying categories as slices of a circle. It effectively illustrates percentage distributions, making it easy to compare different segments. In this analysis, a pie chart is used to show how total advertising expenditure is distributed among TV, Radio, and Newspaper advertisements.

2.2 Histograms

Histograms display the distribution of numerical variables by dividing continuous data into bins and counting the frequency of observations in each bin. They help identify patterns such as skewness, outliers, and the overall shape of the data. Here, histograms are used to analyze the distribution of different features, revealing where data points are concentrated.

2.3 Box Plots

A box plot provides a visual summary of a dataset's distribution using quartiles. It represents key statistics, including the minimum (Q0), first quartile (Q1), median (Q2), third quartile (Q3), and maximum (Q4). These plots are particularly useful for detecting outliers and understanding variability. In this study, box plots are used to examine the distribution of individual advertising expenditures and total advertising costs.

2.4 Violin Plots

A violin plot combines a box plot with a kernel density estimate to illustrate data distribution and variability. It provides insights into where data points are concentrated. Here, violin plots are employed to analyze the density distribution of different features in the dataset.

2.5 Scatter Plots

A scatter plot visualizes the relationship between two or more variables. In particular, a 3D scatter plot is used to explore interactions between three variables, such as TV, Radio, and Sales. By mapping data points in three dimensions (X, Y, and Z), trends and dependencies between features can be identified. In this analysis, scatter plots help examine how advertising expenditures influence sales.

2.6 Q-Q Plots and Normality Tests

A quantile-quantile (Q-Q) plot compares dataset quantiles to a theoretical distribution, typically normal, to assess normality. If data points align closely with a straight line, the data follows a normal distribution; deviations indicate skewness, heavy tails, or outliers. Normality tests such as the Kolmogorov–Smirnov test (for $n \geq 50$) or the Shapiro–Wilk test (for $n < 50$) further confirm normality assumptions. If the data is not normally distributed, transformations or non-parametric tests may be applied. However, since this analysis focuses on continuous data without paired comparisons, correlation metrics are used instead of statistical significance tests.

2.7 Correlation Analysis

Correlation measures the strength and direction of relationships between variables. Spearman's correlation is particularly useful for non-normally distributed or ordinal data, whereas Pearson's correlation assumes normality. In this study, Spearman's correlation is used, as some features do not follow a normal distribution.

2.8 Bar Graph

A bar graph represents categorical data using rectangular bars, where the length of each bar corresponds to its value. It is useful for comparing different categories. In this analysis, bar graphs illustrate advertising expenditures across different sales categories (low and high sales).

2.9 Line of Best Fit

The line of best fit is a regression-based technique that models the relationship between variables while minimizing prediction error. It provides insight into which features most strongly influence the target variable. Here, it is used in scatter plots to highlight trends in the data.

2.10 Advertisement-to-Sales Ratio

The Advertisement-to-Sales (A/S) Ratio is a key metric for assessing the efficiency of advertising investments in generating revenue. It is calculated as follows:

$$\text{A/S Ratio} = \frac{\text{Total Advertising Expenditure}}{\text{Sales Revenue}} \tag{1}$$

A lower ratio suggests higher efficiency, whereas a higher ratio indicates greater spending relative to revenue. Industry-specific benchmarks provide meaningful comparisons [6].

2.11 Descriptive Statistics

Descriptive statistics summarize and organize data through measures of central tendency (mean, median, mode) and dispersion (range, variance, standard deviation). Initially, a statistical summary of the dataset is presented to provide an overview of key features. This is followed by visualizations that offer deeper insights, allowing us to draw conclusions and make informed interpretations of the data.

3 Descriptive Statistics

Metric	TV Cost (1000$)	Radio Cost (1000$)	Newspaper Cost (1000$)	Sales (M$)
Mean	147.042	23.264	30.554	14.022
Median	149.75	22.9	25.75	12.9
Std Dev	85.639	14.810	21.724	5.204
Min	0.7	0.0	0.3	1.6
Max	296.4	49.6	114.0	27.0

Table 1: Statistical Summary of Advertising Dataset

4 Visualization

4.1 Pie Chart

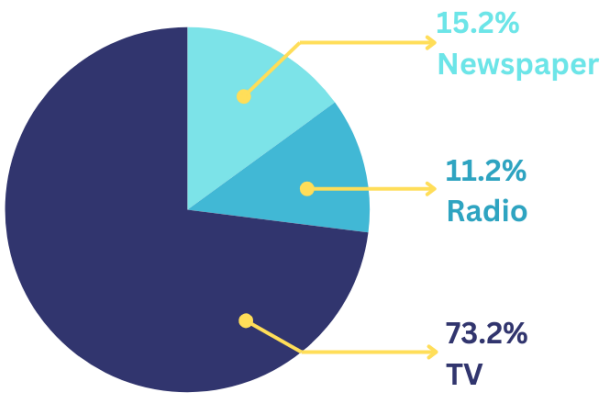


Figure 2: Pie Chart displaying proportion of expenditure in different ad types.

Observations

On average, TV ads receive the highest investment at 73.2%, followed by newspapers at 15.2%, and radio at 11.2%, indicating a strong preference for TV advertising across all data points.

4.2 Histograms

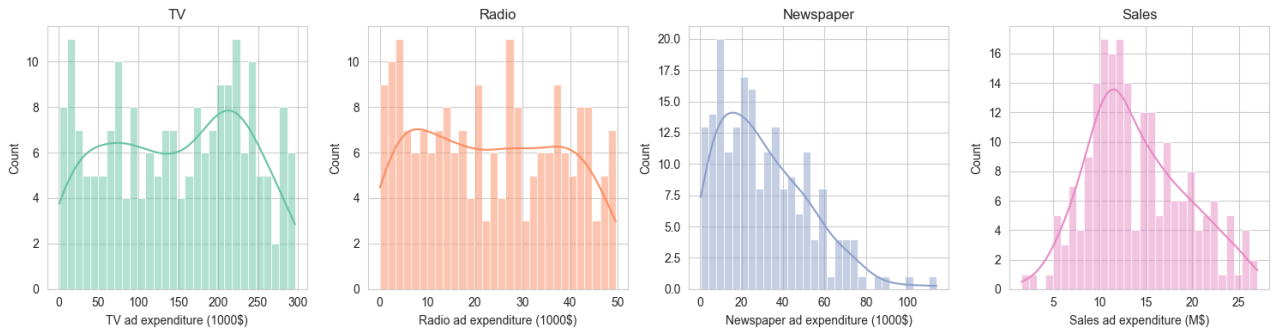


Figure 3: Histogram of all the columns in the dataset.

Observations

1. **TV Ads**: Highest & most consistent investment; preferred medium with uniform spending distribution.
2. **Radio/Newspaper**: Newspaper budgets are right-skewed (concentrated in low range).
3. **Spending Patterns**: TV/Radio evenly distributed; Newspaper/Sales are right-skewed.

4.3 Box + Violin Plots

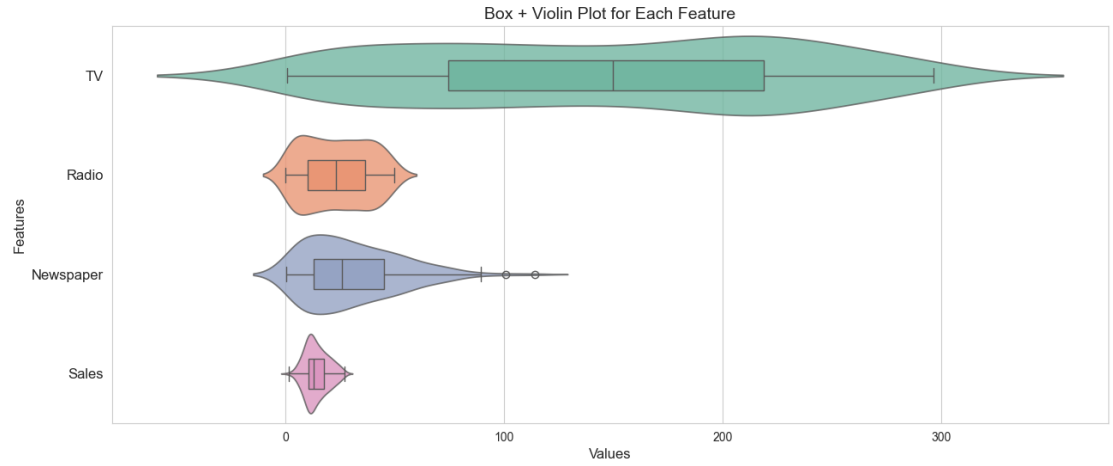


Figure 4: Violin plot (with an overlaid Box plot) for all the features.

Note: Sales is in denomination of Million $ while other features are in denomination 1000$.

Observations

1. **TV Ads**: Widely distributed with a large IQR, indicating varied spending across companies.
2. **Radio Ads**: Concentrated in the lower range, with most companies allocating moderate budgets.
3. **Newspaper Ads**: Right-skewed, with most spending in the lower range and a few high-budget outliers.
4. **Sales**: Mostly moderate revenue, with a short IQR suggesting less variation in sales figures.

4.4 Scatter Plots and Line of best fit

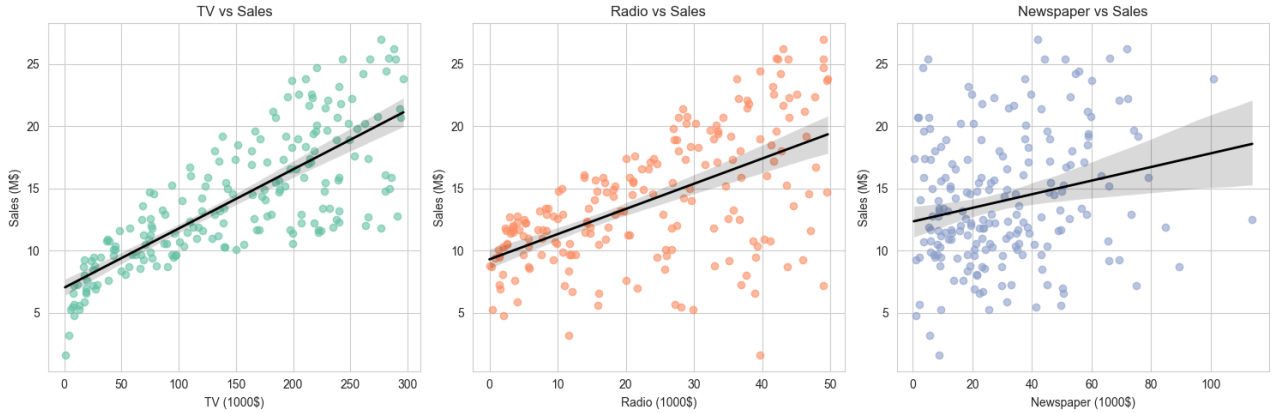


Figure 5: Scatter plot with overlaid line of best fit.

Note: The gray region in the plots denote the 95% confidence interval (C.I.) for the line of best fit.

Observations

1. **TV Ads**: TV ads have the strongest and most direct impact on sales, making them the most effective advertising channel.
2. **Radio/Newspaper**: Radio ads contribute to sales but with more variability, indicating they may not always lead to consistent revenue growth.
3. **Spending Patterns**: Newspaper ads show the weakest impact on sales, suggesting they may not be a reliable investment for driving revenue.

4.5 Total Expenditure Visualization

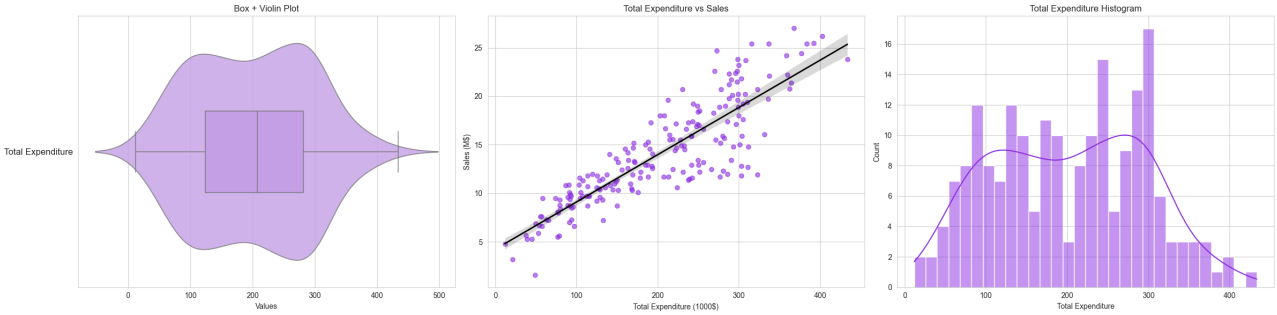


Figure 6: Visualization of Total Advertisement Expenditure using box/violin plot, scatter plot (vs Sales) and Histogram.

Observations

1. The distribution of total advertisement expenditure is spread across a wide range.
2. Total expenditure shows a strong positive correlation with sales, confirming that higher ad spending boosts sales.
3. The histogram reveals a slightly right-skewed distribution of total expenditure.

4.6 Q-Q Plot

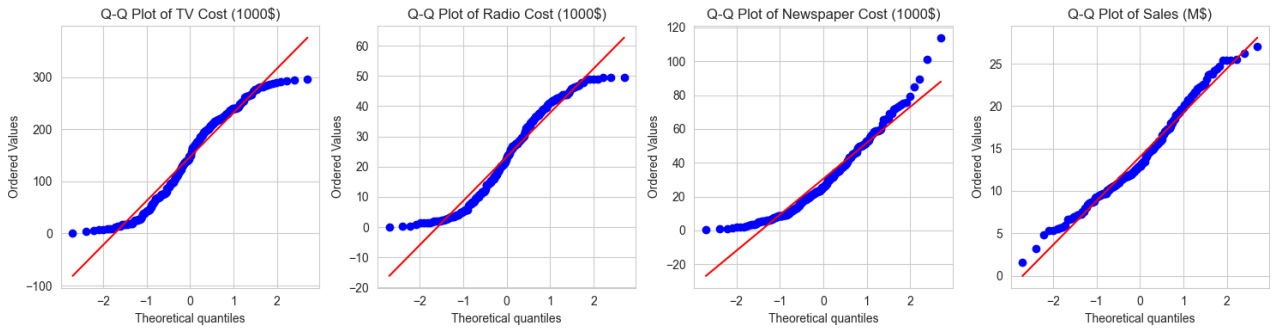


Figure 7: Q-Q (Quantile-Quantile) plot for each feature and target variable.

Note: The red line depicts normally distributed data.

Observations

1. The Q-Q plots for TV, Radio, and Newspaper costs show deviations at the tails (heavy left tails and light right tails), indicating right-skewed distributions with some high-value outliers.
2. The Q-Q plot for sales follows the red line more closely, suggesting sales data is nearly normally distributed, with minor deviations at the extremes.
3. The curvature and upward deviation in the higher quantiles indicate potential outliers in advertising expenditures, especially for newspaper costs, which show a more pronounced deviation.

4.7 Normality Test

```
For TV:
KS Statistic: 0.087
P-value: 0.0911
Fail to reject the null hypothesis: Sample follows a normal distribution.

For Radio:
KS Statistic: 0.084
P-value: 0.1126
Fail to reject the null hypothesis: Sample follows a normal distribution.

For Newspaper:
KS Statistic: 0.0985
P-value: 0.0384
Reject the null hypothesis: Sample does NOT follow a normal distribution.

For Sales:
KS Statistic: 0.0952
P-value: 0.0499
Reject the null hypothesis: Sample does NOT follow a normal distribution.
```

Figure 8: Kolmogorov-Smirnov test of all the features.

Note: I used Kolmogorov-Smirnov test instead of Shapiro-Wilk as the number of rows in my data (=200) $\geq$ 50.

Observations

We can see that Ad expenditure for TV and Radio follows normal distribution whereas expenditure on Newpaper and overall sales do not.

4.8 Correlations

Note: Since we have continuous data rather than groups we cannot use any of the group tests (like t-test, Mann-Whitney U test, etc.) and since our sales variable is not normally distributed we will use the spearman's rank correlation to test for the correlation between variables.

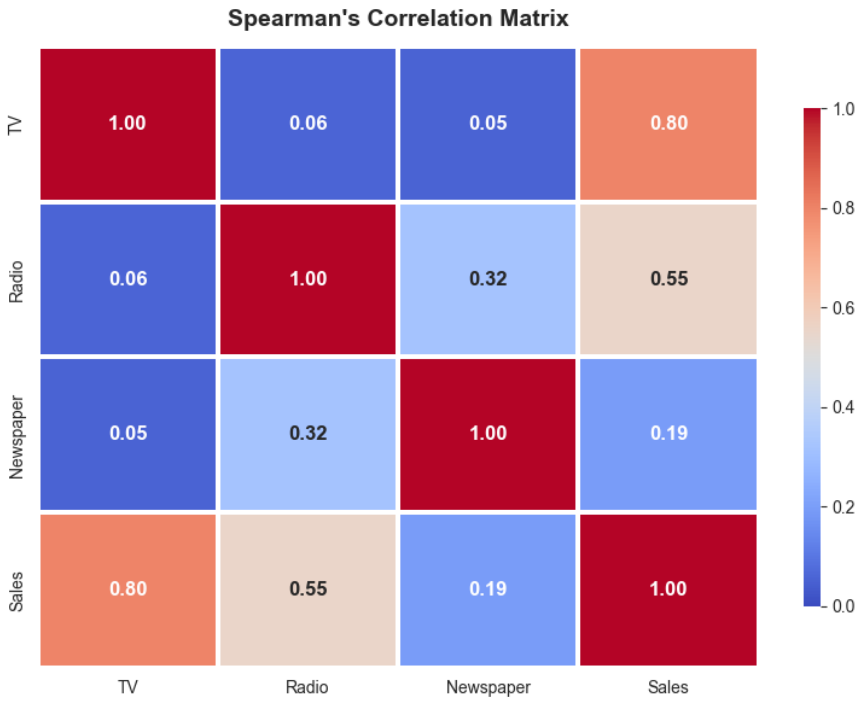


Figure 9: Spearman's Correlation Matrix for all the different features.

Observations

1. TV ad spending has the strongest correlation with sales (0.80), followed by radio (0.55), while newspaper has the weakest correlation (0.19).
2. There is minimal correlation between TV and other ad channels, indicating independent budget allocations.
3. Newspaper and radio have a mild correlation (0.32), possibly suggesting overlapping audience targeting.

4.9 Advertisement to Sales Ratio

Observations

Fig. 10 shows most companies have an A2S ratio that falls within the mid-range with an average around 0.014, with only a few exhibiting extreme values. Additionally, the Q-Q plot suggests that the data follows a normal distribution.

5 Question Answers

Q1 Which platform yields the highest return on investment for advertising campaigns?

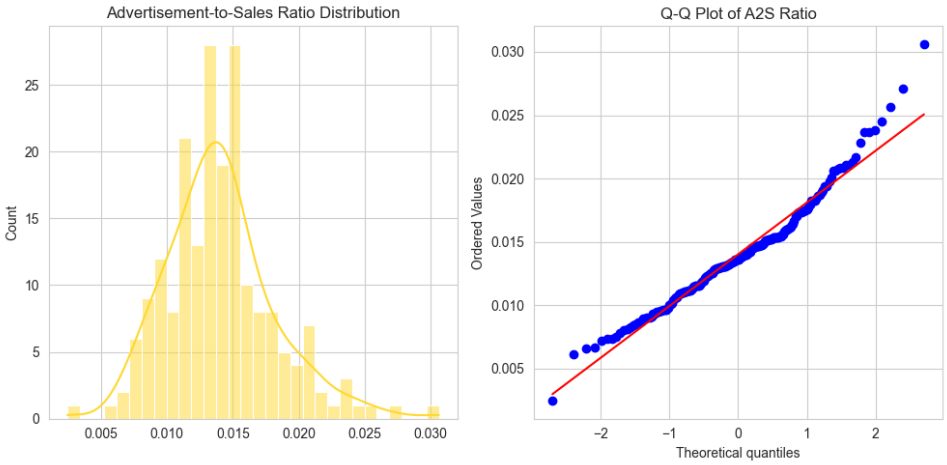


Figure 10: Advertisement to Sales ratio values visualised in histogram and Q-Q Plot.

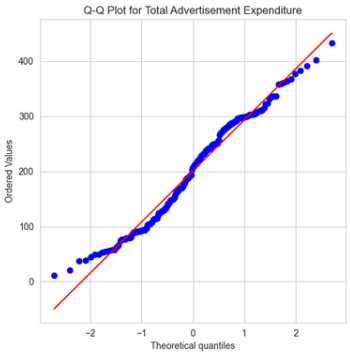
A From the correlation heatmap 9 and scatter plots 5, TV advertising shows the strongest positive correlation with sales. The regression line in scatter plots 5 also suggests that TV ad spend has the most direct impact on sales compared to Radio and Newspaper. Hence we can say that **TV advertising provides the highest return on investment (ROI) among the three platforms.**

Q2 Is there a general trend between TV advertising spend and sales?

A The scatter plot (TV vs. Sales) with regression line 5 shows a clear upward trend, meaning higher TV ad spend generally leads to higher sales. The strong correlation coefficient (0.80) 9 between expenditure on TV Ads and Sales also supports this relationship. Hence we can say that **there is a strong positive trend between TV advertising spend and sales.**

Q3 What is the relationship between ad expenditure and sales?

A As can be seen from the scatter plot of total expenditure on advertisements vs revenue in sales in figure 6 that **they have a strong positive linear relationship**. And to confirm this we first did the normality check On calculating their correlation value turned out to be **0.87** (Fig. 11) hence confirming the linear positive correlation.




```
KS Statistic: 0.0702
P-value: 0.2657
Fail to reject the null hypothesis: Total Expenditure follows a normal
distribution.

Spearman's Rank Correlation between Total Ad Expenditure and Sales:  0.8771
```

Figure 11: Q-Q plot for total expenditure on advertisements along with significance test statistics.

Q4 How do the average advertising spends compare when sales are high versus when sales are low?

A With reference to the scatter plot between total expenditure on Ads and revenue from Sales 9 we can clearly see that **when the sales are low the average advertising spends are also low and vice versa**. To verify this I calculated the average total expenditure on advertisements and the results are as follows:

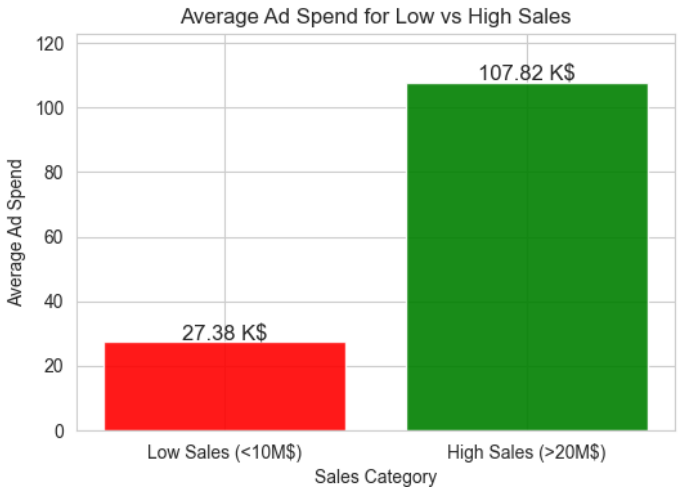


Figure 12: Bar Graph showing the difference between expenditure on Advertisements when the sales were low and high respectively.

Please note that sales below $\$10,000,000$ are considered low and sales above $\$20,000,000$ are considered high. Hence they kind of show a direct proportionality.

6 Conclusion

The analysis of advertising expenditures across TV, Radio, and Newspaper has provided critical insights into their effectiveness in driving sales revenue. The descriptive statistics indicate that TV advertising receives the highest investment and has the strongest impact on sales, with a high correlation of 0.80. Radio advertising follows, showing moderate effectiveness, while Newspaper advertisements contribute the least to revenue generation. Various statistical methods, including visualization techniques, correlation analysis, and normality tests, confirm that TV is the most influential medium for sales growth.

From the scatter plots and correlation analysis, we observe a strong positive linear relationship between total advertising expenditure and sales. This confirms that increased ad spending generally leads to higher sales. However, spending patterns vary, with Newspaper expenditures being right-skewed, indicating a concentration of lower-budget campaigns with a few high-value outliers. Additionally, the normality tests reveal that TV and Radio expenditures follow a normal distribution, whereas Newspaper spending and sales do not, necessitating the use of non-parametric methods for further analysis.

The effectiveness of advertising spending is not uniform across different sales levels. Companies with higher sales tend to invest more heavily in advertisements, particularly in TV campaigns. The analysis of sales quartiles and spending patterns will provide further insights into how ad effectiveness varies across different revenue segments. Additionally, by evaluating the diminishing returns of advertising for each channel, we can determine the optimal spending threshold beyond which additional investments may yield limited benefits.

To extend this study and deepen the insights, the following key research questions will be explored in the next phase of analysis:

- What is the optimal allocation of advertising budget across TV, Radio, and Newspaper to maximize sales?
- How does advertising effectiveness differ across spending level categories (low, medium, high) for each channel?
- At what spending levels do returns begin to diminish for each advertising channel?
- What interaction effects exist between different advertising channels and how do they influence sales?

To address these questions, advanced statistical techniques will be applied, such as **Multiple Linear Regression, ANOVA (One-Way & Factorial, Generalized Linear Models (GLMs), Interaction Analysis**.

By incorporating these methodologies, we aim to refine our understanding of advertising efficiency and develop data-driven strategies to optimize marketing investments. The next report will provide an in-depth analysis addressing these questions, ultimately guiding businesses toward more effective advertising decisions.

References

[1] Wikipedia. *History of Advertising*. https://en.wikipedia.org/wiki/History_of_advertising/. Accessed: 20-Feb-2025.

[2] Power Direct Marketing. *A Brief History of Print Advertising*. https://powerdirect.net/history-print-advertising/. Accessed: 20-Feb-2025.

[3] Spotify Editorial Team. *The history of radio advertising and the state of audio today*. https://ads.spotify.com/en-US/news-and-insights/history-of-radio-advertising/. Accessed: 20-Feb-2025.

[4] Wikipedia. *Television advertisement*. https://en.wikipedia.org/wiki/Television_advertisement/. Accessed: 20-Feb-2025.

[5] Wikipedia. *Digital Marketing*. https://en.wikipedia.org/wiki/Digital_marketing/. Accessed: 20-Feb-2025.

[6] Will Kenton. *Advertising-To-Sales Ratio: Overview, Examples, How to Read It*. https://www.investopedia.com/terms/a/advertising-to-sales-ratio.asp/. Accessed: 20-Feb-2025.