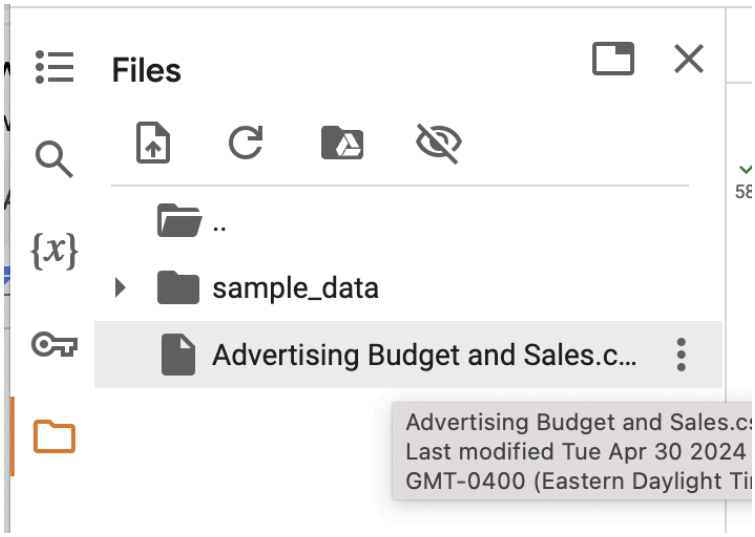
How do I run the code?

Simple!

- Drag the file attached named “Advertising Budget and Sales.csv” into this file box then click run!
- If the code does not run because you need to change the name of the file before you drag it in, if possible please click the 3 dots, copy path and paste the link in the line filepath = “” in the main function.
- It should however run, once this exact file is placed where it is shown in the screenshot.
- Thank you!!!



My program is designed to analyze and predict the impact of advertising spend on sales across multiple channels, including TV, radio, and newspaper. Particularly, it will allow you to determine the amount you should invest into each channel for maximum . Initially, it loads advertising data from a CSV file and performs preliminary cleaning to remove any instances with missing values, ensuring the integrity of the dataset. The program then computes descriptive statistics to give an overview of the central tendencies, dispersions, and distributions of the advertising spends and sales. It uses histograms to visualize the distribution of expenditures across different media, providing initial insights into the data structure and potential outliers.

Further, the program calculates correlation matrices to explore the relationships between different advertising channels and sales, which informs the feature selection for model building. It creates interaction features to examine potential synergistic effects between different advertising channels. Using this prepared dataset, the program trains multiple regression models, including Linear

Regression, Decision Tree, and Support Vector Regression, to predict sales. It evaluates these models based on R-squared and RMSE metrics to select the best-performing model. Once the optimal model is identified, the program uses it to simulate the impact of varying advertising budgets on sales and provides a recommendation for budget allocation. Additionally, the program employs K-Means clustering to identify patterns in budget allocation and uses Sankey diagrams to visually represent the flow of spending to sales. Finally, it enables user interaction by allowing users to input their advertising budgets and predicting sales outcomes based on the model, making the program a robust tool for strategic planning and optimization in advertising.

The **computation** aspect of this project was realized through the development of functions designed to handle data manipulation and modeling. The read_data function streamlines the import process, allowing for a quick transition to data processing with clean_data, which tidies the dataset for analysis. Computationally intensive tasks, such as creating interaction features through create_interaction_features and executing machine learning models in train_multiple_models, highlight the program's capability to handle complex calculations. These functions showcase the application's ability to not only handle data but also in preparing it for analytical consumption and predictive modeling, satisfying the computational requirement.

For the **statistics** requirement, the program uses functions like summary_statistics to provide descriptive insights into the data, while compute_correlation_matrix and perform_t_test offer inferential statistical analysis. The former measures the strength and direction of relationships between variables, and the latter tests hypotheses about these relationships. By integrating the k_means_clustering function, the program extends its statistical capacity to segment the market, identifying patterns and groups within the advertising data that are not immediately visible, thereby providing a nuanced statistical understanding of the dataset.

Visual representation of data is achieved through multiple tailored functions. The plot_distribution function visually summarizes the spread and central tendency of each advertising channel, providing immediate visual insights. plot_correlation_heatmap offers a heat map of correlation coefficients, which is an effective way to visualize complex relationships. Advanced visualizations are employed in plot_cluster_results, revealing data segmentation, and plot_dynamic_visualization, illustrating the regression line and potential patterns in the data. The inclusion of plot_sankey_diagram and plot_residuals further showcases the program's comprehensive visualization capabilities, providing a holistic view of both the distribution of spends and model performance, fulfilling the data visualization requirement.

In this project, each function within the Python script serves a specific purpose to contribute effectively toward the overarching goal of optimizing advertising budgets to maximize sales performance.

The data analysis process begins with the ReadData(filepath) function, which loads data from the specified CSV file into a pandas DataFrame. This function sets the stage for all subsequent data manipulations and analysis. Following the input of data, the CleanData(dataframe) function

enhances data quality by removing rows with any missing values. Ensuring the dataset is complete and accurate is crucial, as missing data can lead to biased or incorrect statistical analyses and predictive modeling outcomes. This also served as a **limitation** as perhaps a more representative solution would have been to replace empty values with the mean or median however this is also not fully representative.

To provide an initial understanding of the dataset, SummaryStatistics(dataframe) generates descriptive statistics that offer insights into the data's central tendencies and variability. This function lays the foundational knowledge required for informed decision-making in later stages of the analysis, giving a quick overview of potential anomalies or patterns that may require further investigation.

Visualization functions play a critical role in making the data accessible and understandable. PlotDistribution(dataframe, column, plot_type) allows users to visualize the distribution of a specified column using histograms or boxplots, aiding in the identification of outliers and understanding the spread of the data. This was particularly useful for me in seeing a summary of the data as it showed how much money was spent on each avenue. It could have been improved by checking assumptions of normality or uniformity in the data distribution, which is critical for many statistical tests and models. Similarly, additional functions like plot_each_visualization(dataframe) and plot_feature_importance(model, feature_names) dove deeper into specific channels' impacts and the importance of various features in the models, respectively. These visualizations are instrumental in dissecting the contributions of different advertising channels and understanding which attributes most strongly influence sales predictions.

To visualize the clustering of data, KMeansClustering(dataframe, feature1, feature2, max_clusters) and PlotClusters(dataframe, feature1, feature2, labels) are used. These functions apply and then plot the results of KMeans clustering, helping to identify natural groupings within the data that can inform targeted marketing strategies.

The ComputeCorrelationMatrix(dataframe) function calculates Pearson, Kendall, and Spearman correlation coefficients for all pairwise relationships within the dataset. Understanding these correlations is crucial for feature selection in predictive modeling as it helps identify potential predictors for the dependent variable. The PlotCorrelationHeatmap(correlation_matrix, method) function then visually represents these relationships through heatmaps, allowing for easy identification of highly correlated variables, which might influence the model development process by indicating redundancy or potential multicollinearity issues.

Further into the analysis, the CreateInteractionFeatures(dataframe) function enhances the dataset by calculating interactions between selected advertising budget columns. This method can unveil complex relationships and synergies between different advertising channels, which might not be captured by simple additive models. The new features created by this function can significantly contribute to the nuanced models that better predict sales outcomes based on advertising spends.

Modeling functions such as TrainMultipleModels(dataframe, dependent_var, feature_list) and SelectBestModel(model_performance) are central to identifying the best predictive model. These functions train multiple regression models, compare their performance using metrics like RMSE and R-squared, and select the best performer for making predictions. This rigorous evaluation ensures that the selected model is the most suitable for predicting sales based on the given advertising budgets. Further, plot_residuals(model, X, y) helps evaluate model performance by visualizing residuals against predicted values, offering insights into any systematic errors that the model may be producing.

Simulation and advanced visualization functions like SimulateAdvertisingImpact(dataframe, best_model, feature_list) and PlotSankeyDiagram(dataframe) provide strategic insights into the allocation of advertising budgets and their impact on sales. These functions allow stakeholders to visualize potential outcomes of different budget scenarios and understand the flow of advertising spend towards sales in a visually intuitive manner.

The plot_residuals function plot scatter plot of residuals vs the predicted values from the best regression model. The model_validation_metrics function then runs more tests to calculate performance metrics of the model. Lastly, the allocate_optimal_budget function allows the user to enter their budget and tells the user what they recommend they should invest into each avenue and then what the prediction of sales would be given this investment or with a different allocation of their choosing.

I did experience some issues when working with my dataset. My first dataset was not working as expected and as such, I had to pivot and find a new dataset to use. My new plan was to do all the analysis on both datasets and compare them however the first dataset still posed challenges and as such, I stuck with the new dataset. In my analysis of advertising data, several assumptions and simplifications were necessary to streamline the process and focus on the most impactful insights. First, I assumed that the dataset was largely complete and representative of typical advertising scenarios, which allowed me to proceed without extensive data augmentation or complex handling of outliers.

A significant simplification was in the handling of missing data. I chose to remove any rows with missing values outright, which, while expedient, can potentially bias the dataset if the missingness is not random. This approach assumes that the missing data is minimal and does not contain patterns that could influence the analysis outcomes. Also, my combination function assumed linearity as a simplification even though it was not the model that was best suited. However, it was very close to the model which was chosen. Ideally, this function would use the best model to find its combinations.

Additionally, I fell into an issue due to the structure of my dataset. Since the format is such that there are many channels that contribute to the output, it is difficult to pinpoint regression in individual categories as that category may not be truly responsible for the outcome since there are also

contributions from 5 other channels at the same time. There was also an issue due to the actual data since it was not equally represented across all of the channels.

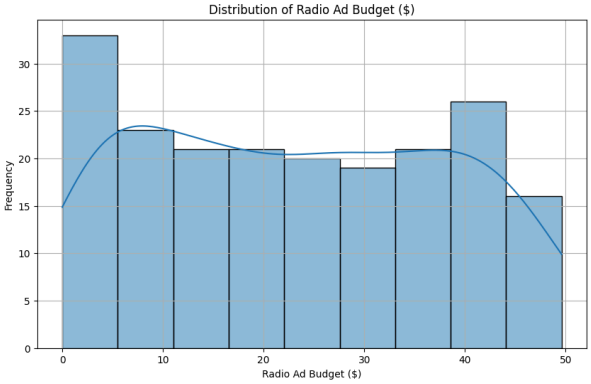
Regarding the interaction features, I simplified the analysis by only considering pairwise interactions between advertising channels. This decision was made under the assumption that higher-order interactions (involving three or more channels) would not significantly alter the insights gained from the pairwise interactions. This reduces computational complexity and focuses the analysis on the most direct relationships.Additionally, I planned to do a lot of analysis with the create_interaction_feature but it did not end up making sense to do in the scheme of the project.

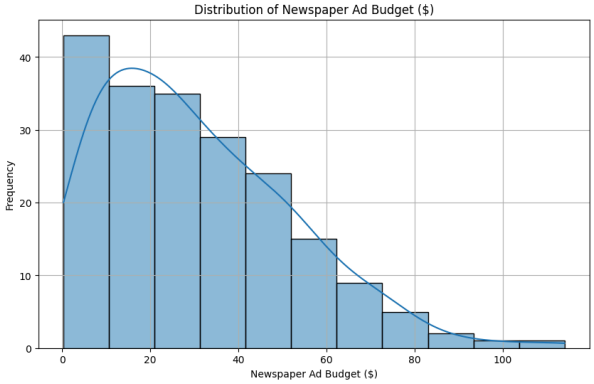
In the statistical modeling phase, I employed linear models which assume a linear relationship between the advertising spend and sales outcomes. This simplification ignores potential nonlinearities or more complex dynamics in how advertising efforts convert into sales. However, it made the analysis more tractable and the results easier to interpret. Also, I changed the combination function which was supposed to predict sales based on a combination of advertising methods however, I used a linear method which would not have been accurate considering that the best model was the Decision Tree so in the end, this function was removed.

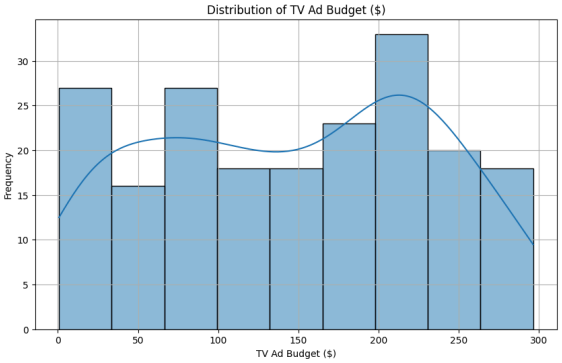
Lastly, for the K-Means clustering used to segment the data, I assumed that the clusters would capture meaningful market segments based on spending patterns. This assumes that the standardization of data (using z-scores) adequately normalizes the scale of spending across different channels, which might not always reflect true market conditions where different scales could have different implications on customer behavior. Unfortunately, my K-Means clustering did not provide very useful information which could stem from the fact that the data in my dataset was stimulated and does not reflect real data and real patterns.

Lastly, looking to the future with this project, there are several ways that it could be advanced/improved. I think the most important enhancement would be to implement time data. This would allow us to see how much sales are generated after x amount of time for each media channel. Also, to Implement time series analysis to account for trends and seasonality in advertising spend and sales data. This could help in forecasting future trends and determining optimal times for ad campaigns. Also, to explore more advanced machine learning models, such as ensemble methods (Random Forests, Gradient Boosting Machines) or neural networks, to potentially improve predictive accuracy. I also think it would be better to have more user interaction by developing an interactive web dashboard where users can visualize data, input parameters, and view predictions and recommendations in real-time, enhancing the user experience and accessibility.

Key features in plots are shown on the next page







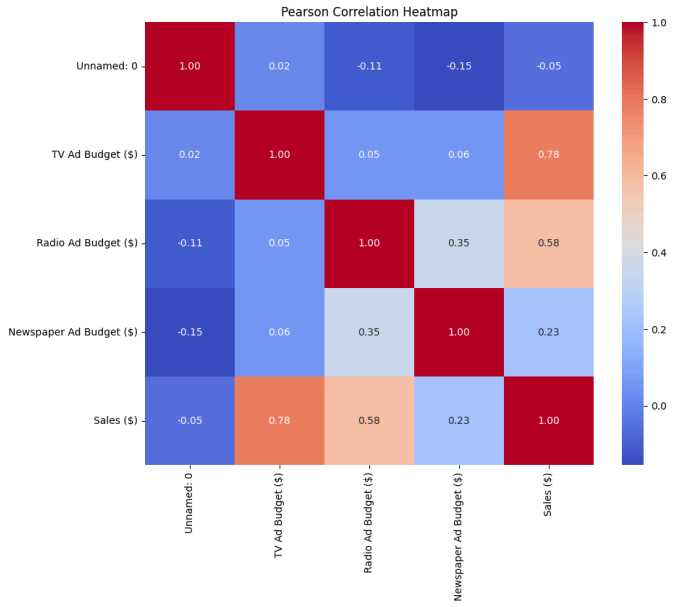
These histograms were plotted as they are representative of the data which will be worked with throughout the project. It is apparent that while different values of TV and radio budgets are tested throughout the dataset (Apparent as their curve is relatively stable throughout), only small amounts of larger budgets have been made for the newspaper Ads in the dataset. This is not very representative and is definitely going to be a limitation since the models will not have the information of what happens if for example, a large amount of money is invested into newspaper ads. In my opinion, the data would be more useful if equal amounts of high and low investments were made in each and the coordinating outcomes of such a result. Hence, the data does seem a bit biased. However, this does make sense in the real world as one would expect a lot more money to be invested into TV or Radio ads compared to newspapers especially with the decline of physical newspaper sales nowadays. Certainly, the type of product being sold would also have an impact on where would be the most impactful avenue to put funds in for advertisements. Additionally, it should be of note that the data in this dataset is unrealistic as seen by the very low average in sales for higher averages in advertising. This is a poor return on investment and with real data this trend will definitely not be seen or else advertising does not make much sense.

The graph below shows the measures of tendencies calculated. These line up with what is seen on the histograms.

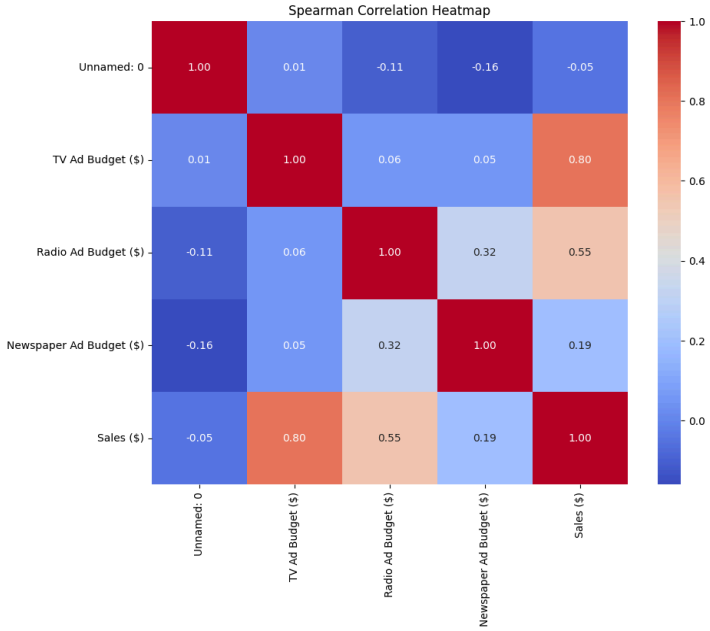
Summary Statistics Calculated in the code

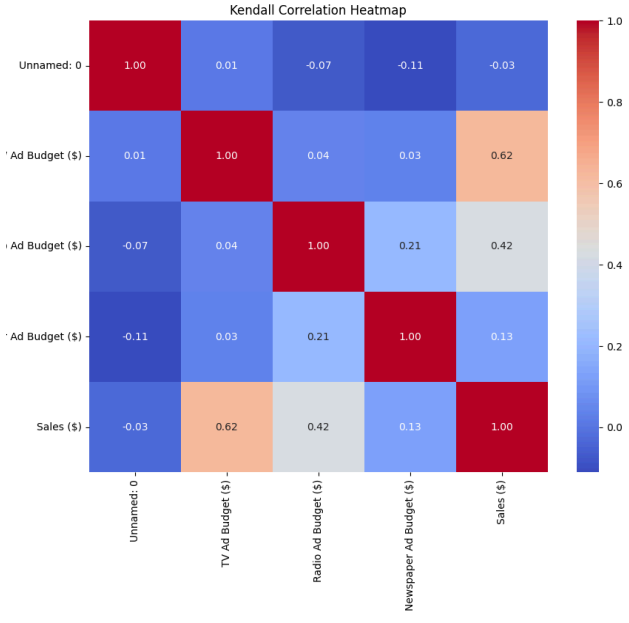
	count	mean	std	min	25%	50%	75%	max
TV Ad Budget ($)	200.0	147.0425	85.854236	0.7	74.375	149.75	218.825	296.4
Radio Ad Budget ($)	200.0	23.2640	14.846809	0.0	9.975	22.90	36.525	49.6
Newspaper Ad Budget ($)	200.0	30.5540	21.778621	0.3	12.750	25.75	45.100	114.0
Sales ($)	200.0	14.0225	5.217457	1.6	10.375	12.90	17.400	27.0

Next, correlation matrices were calculated using three different methods (Pearson, Kendall and Spearman)

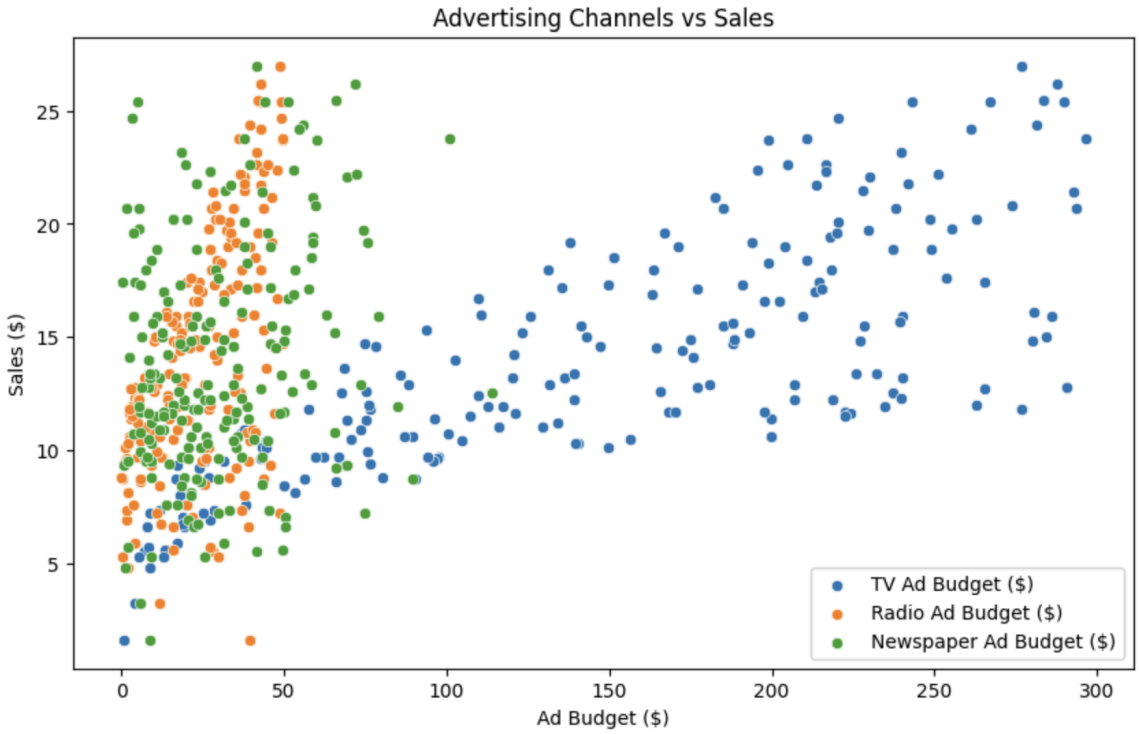


It can be seen in all three heatmaps, that the TV Ad budget had the greatest correlation with sales with a coefficient of 0.78, 0.8 and 0.62, followed by radio with 0.58,0.55 and 0.42, and then newspaper with 0.23,0.19 and 0.13 in the Pearson, Spearman and Kendall Matrices respectively. Other moderately strong correlations include that of the radio ad budget and newspaper ad budget with around 0.3 coefficient taking into consideration all three models.





```
Plotting all channels vs sales on the same axis...
```

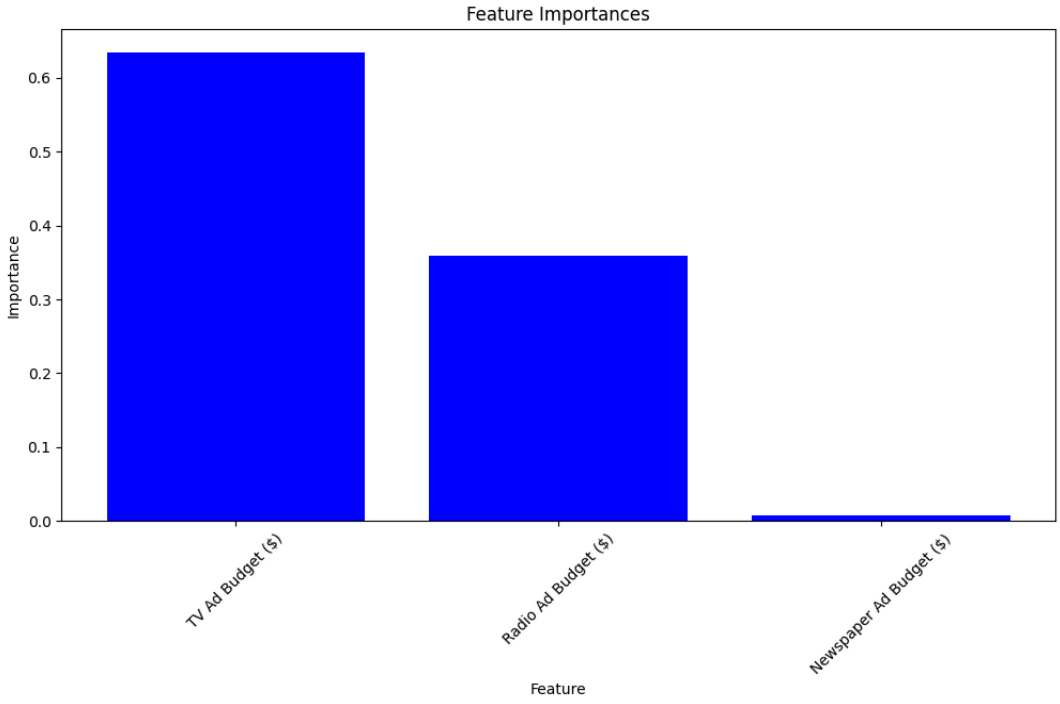


The graph above shows each Ad budget against Sales and is really useful in seeing how they compare to each other when looked at side by side. It is clear that there have been much bigger investments in the TV Ad Budget. Additionally since sales reflects a certain budget of all three combined, it may be difficult to determine what sales is a result of from this diagram alone.

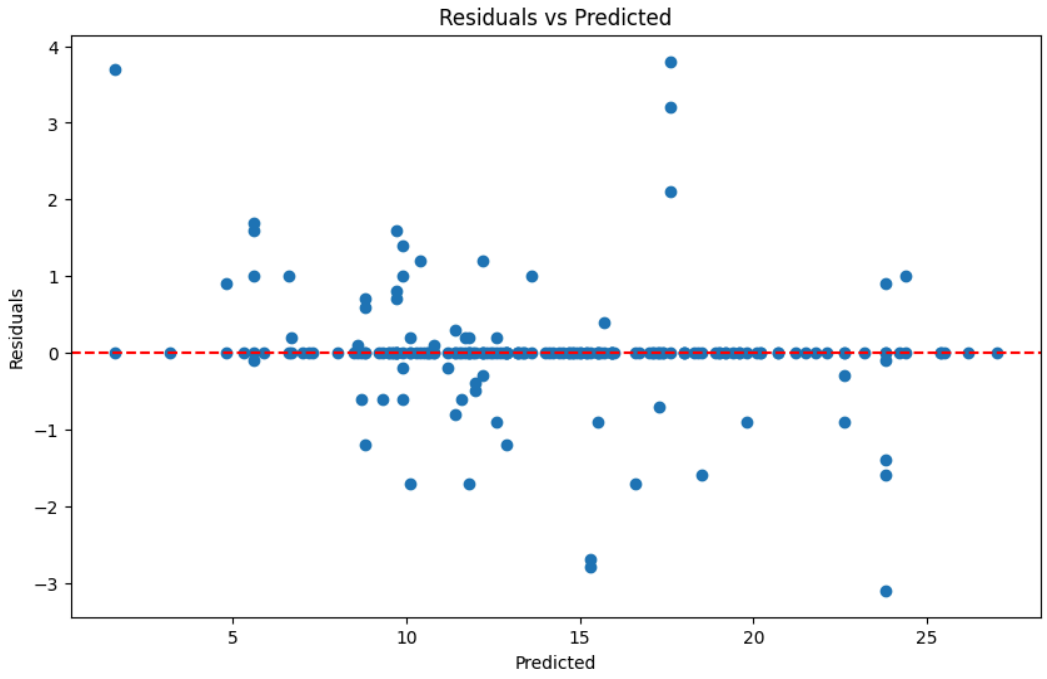
The results from fitting the data using machine learning models is shown below:

Different Models		
Model	R^2	RMSE
Linear Regression	0.86095	1.9485372043446387
Decision Tree Regressor	0.94422797	1.2340313340160105
SVR	0.840999	2.089499680388078

The **Decision Tree Regressor** was chosen as the most suitable model as it presented the lowest mean square error of the 3 methods.



This diagram shows that TV ad Budget has the most importance with a value of 0.6, radio with 0.4 and newspaper close to 0. This is consistent with what the other graphs have shown thus far. It has to again be noted that this is related to the very low values of the newspaper budget presented in the data.

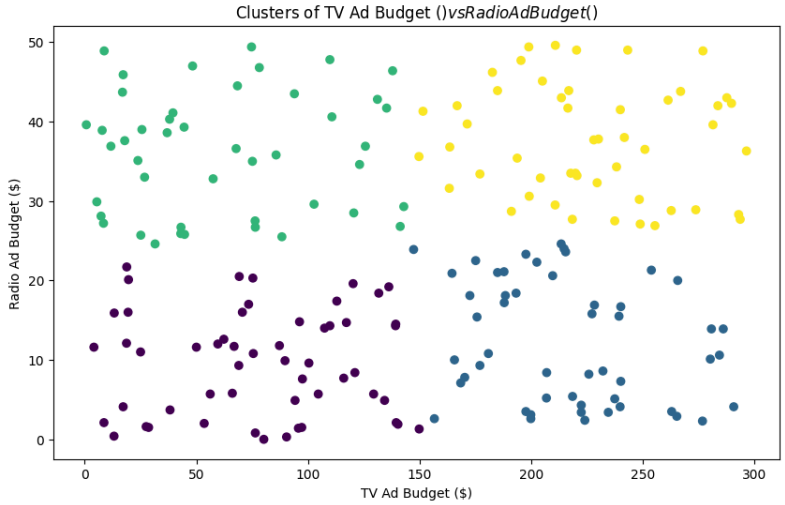


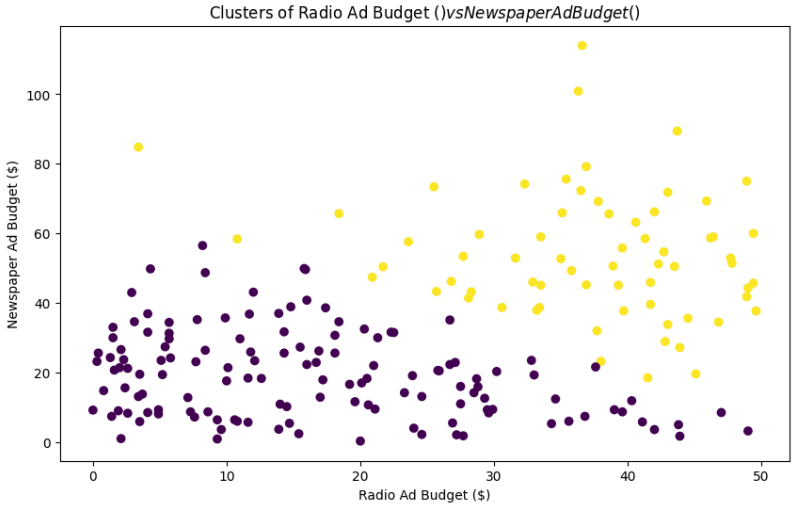
Testing Model	
R-squared	0.9831331934534749
RMSE	0.6759067983087611
MAE	0.2945000000000001
MAPE	0.026187927783224228

Scatter Pattern: Ideally, residuals should be randomly distributed around the horizontal line at zero (no pattern). In my plot, the residuals appear reasonably random without clear patterns, which is good as it suggests that the model does not suffer from non-constant variance or missed non-linear relationships.
R-squared (0.9831): This value is very close to 1, which suggests that the model explains approximately 98.31% of the variance in the dependent variable. It indicates a high level of model fit, showing that the model predictions are very consistent with the actual data.
Root Mean Square Error (RMSE) (0.6759): RMSE provides the standard deviation of the residuals. A lower RMSE is better as it indicates that the errors between predicted and actual values are smaller. In this case, an RMSE of about 0.68 suggests that the model predictions deviate from the actual values by roughly 0.68 units on average, which is relatively low.
Mean Absolute Error (MAE) (0.2945): MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. A MAE of 0.2945 suggests that the average error in the predictions is also quite low.
Mean Absolute Percentage Error (MAPE) (0.0262): MAPE expresses accuracy as a percentage. A MAPE of 2.62% is excellent, indicating that the model predictions are, on average, within 2.62% of the actual values

K- means clustering on the different budgets against each other

In the context of this dataset, the K-means tests were not particularly useful. A simplification was made here in that at first I used an algorithm to compress all the features into a z score and then plot with sales to try to see the impact they had on sales. However, this did not prove useful and instead, I simplified to plot the different features against each other.

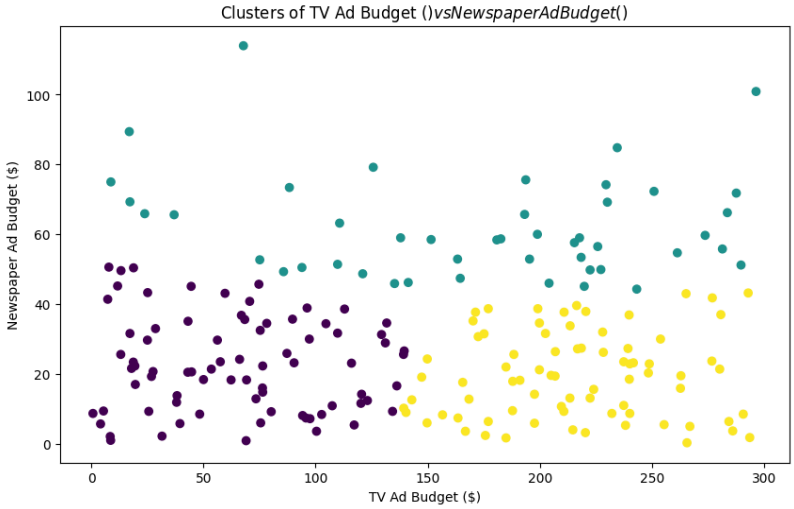




For the first diagram, the data is separated into 4 main clusters. It seems that the data points may each possibly represent different strategies or company types depending on their spending patterns on TV and Radio. Each cluster may suggest different marketing strategies:
For example, one cluster might represent high spenders on TV with moderate spending on Radio, while another represents lower spending on both.
This can help identify which combinations of TV and Radio budgets are common in the dataset, potentially informing how businesses typically allocate their budgets between these two media. Plotting sales next to this could be useful to see how these clusters correspond to sales.

.

In the second, 2 main clusters were observed. This could indicate diverse strategies in the integration of Radio and Newspaper advertising. For instance, some clusters may show a preference for higher Radio budgets over Newspaper, or vice versa.



For the third plot, 3 clusters were observed showing that there may be 3 main categories that these budgets fall into.By analyzing which clusters correlate with higher sales (if sales data is plotted or analyzed in conjunction), businesses could optimize their advertising spend by shifting more budget to the more effective channel.

Simulating the impact of different advertising budgets on product sales…

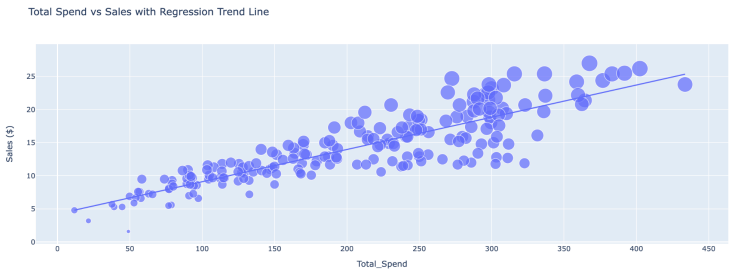
	Simulated_Sales_10%	Simulated_Sales_20%	Simulated_Sales_30%
0	25.5	26.2	27.0
1	10.4	11.8	11.8
2	9.3	9.3	9.3
3	19.6	21.2	23.8
4	14.8	15.2	15.7

	Simulated_Sales_40%	Simulated_Sales_50%	Simulated_Sales_60%
0	27.0	27.0	27.0
1	11.8	15.3	15.3
2	9.3	9.3	9.3
3	23.8	25.4	25.4
4	15.7	15.7	15.7

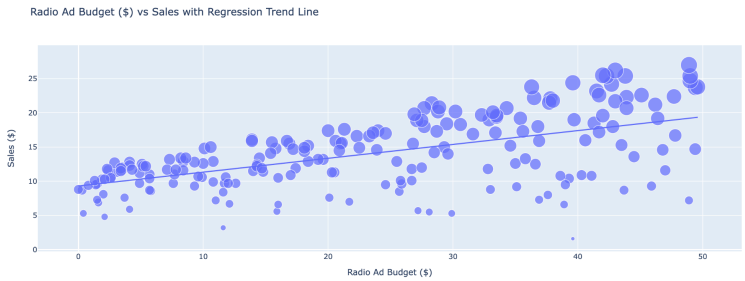
	Simulated_Sales_70%	Simulated_Sales_80%	Simulated_Sales_90%
0	27.0	27.0	27.0
1	15.3	15.3	15.3
2	9.3	9.3	9.3
3	25.4	27.0	27.0
4	15.7	15.7	15.7

Some rows show consistent sales predictions across different budget increases (like rows 0 and 3), suggesting that for these cases, the model predicts that increasing the budget does not significantly impact sales beyond a certain point. Other rows show variability in sales predictions as the budget increases (e.g., rows 1, 2, and 4), which may reflect different sensitivities to budget changes based on the initial conditions or configurations of the advertising spends in these scenarios.

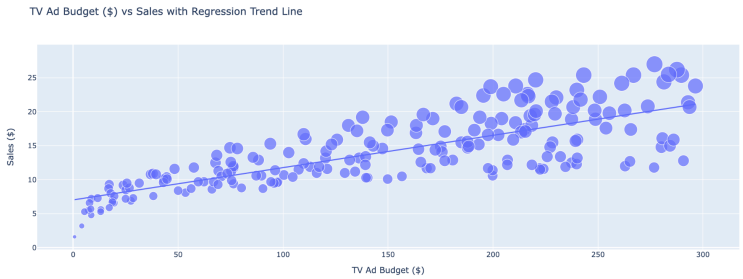
Below are all interactive plots of each Budget against sales as well as the total spent against sales. When these are plots of colab, you can hover over the line and points to get information on each.



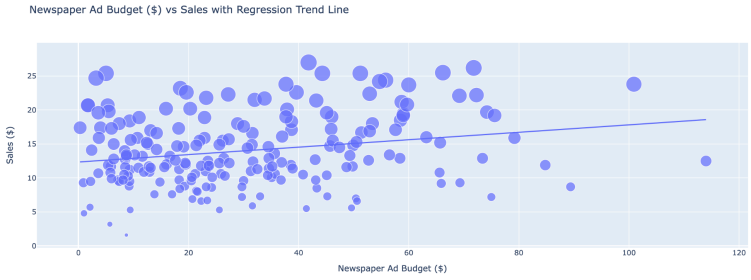
There's a clear positive trend, indicating that as total advertising spend increases, sales also tend to increase. The trend line suggests a linear relationship, although the data points show some variability around this line, indicating that other factors might also influence sales.



The positive slope of the regression line indicates a positive relationship between radio advertising spend and sales. The spread of the data points suggests a moderate correlation, with some variance that isn't explained solely by radio ad spend.



The regression line shows a clear positive trend, indicating a strong relationship between TV ad spend and sales. The data points are somewhat closely packed around the trend line, especially at higher budget levels, suggesting that TV ads might be a highly effective medium for driving sales.



The slight positive slope of the regression line indicates that there is a relationship between newspaper ad spend and sales, but the spread of the data points is quite broad, indicating a weaker correlation than seen with TV and Radio.

This visualization is effective in showing the flow of investment from the advertising budgets to the sales. It can quickly convey the relative weight of each advertising channel in terms of their budget as well as their impact on sales. The width of each band proportionally represents the amount spent on that advertising medium. The fact that the TV Ad Budget is significantly wider than others, indicates a higher spend in that channel. This visualization can help stakeholders understand where money is being invested and how these investments translate into revenue, facilitating more informed budgeting decisions.

The last function asks users to input the total amount they can spend on advertising in general, then suggests how much should be invested into each avenue. The function then allows the user to input these suggested values or edit them a bit to their preference and then predict sales. Eg of an output.

```
Recommended Budget Allocation:
TV Ad Budget ($): $126.92
Radio Ad Budget ($): $71.72
Newspaper Ad Budget ($): $1.36

--- Sales Prediction based on your Advertising Spend ---
Enter your TV Ad Budget ($): 126.92
Enter your Radio Ad Budget ($): 71.72
Enter your Newspaper Ad Budget ($): 1.36
Predicted Sales: $16.00
```

Yay!! Predicted Sales Complete

It must be noted that these small values of sales are unrealistic given the large amount of money put into advertising however it is using the data that was given in the dataset