PROFESSIONAL TRAINING 1

at

Sathyabama Institute of Science and Technology (Deemed to be University)

Submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering

By **P.Larance Kumar**

(Reg.No : 41110953)

DEPARTMENT OF COMPUTER SCIENCE AND

ENGINEERING SCHOOL OF COMPUTING

**SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119, TAMILNADU**

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "**A" by NAAC**

(Established under Section 3 of UGC Act, 1956)

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI– 6001

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

<u>BONAFIDE CERTIFICATE</u>

This is to certify that this Project Report is the bonafide work of **P.LARANCE KUMAR** who carried out the project entitled " **ADVERTISING DATASET**" under my supervision from Jan 2023 to Apr 2023.

Internal Guide

Dr.L.SUJI HELEN

Head of the Department

DR. LAKSHAMANAN M.E., Ph.D.

Submitted for Viva voce Examination held on _____

Internal Examiner   **ExternalExaminer**

2

DECLARATION

I,**P.LARANCE KUMAR** hereby declare that the Project Report entitled **ADVERTISING DATA SET** Done by me under the guidance of **Dr .L.SUJI HELEN, M.E., Ph.D** is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE: 06/10/23

SIGNATURE OF THE CANDIDATE

PLACE: Chennai

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks   to   **Dr. T. Sasikala   M.E., Ph.D.**, **Dean**, School   of Computing ,**Dr. L. Lakshmanan M.E., Ph.D.,** Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide for **Dr .L.SUJI HELEN,** her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful   in many ways for the completion of the project.
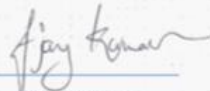
CERTIFICATE OF TRAINING

This certificate is presented to

PENTAMANI LARANCE KUMAR

Register Number: 41110953

From Sathyabama Institute of Science and Technology

Hariharasudhan
Instructor

Ajay Kumar
Director

Certificate ID: 23161/Oct/Cognibot

COGNIBOT
AI meets Industry

The student has successfully completed the 45 hours professional training program on

MACHINE LEARNING

Issue Date: 5th Oct, 2023

Course Duration:
28th July, 2023 to
4th October, 2023

www.cognibot.ml

ABSTRACT

This dataset provides a comprehensive collection of advertising data spanning various industries and mediums. It encompasses a diverse range of campaigns, including digital, print, and broadcast, along with associated metrics such as reach, engagement, and conversion rates. The dataset offers a valuable resource for marketers, analysts, and researchers seeking insights into consumer behavior, campaign effectiveness, and market trends. With a wide array of variables, it enables in-depth exploratory analysis, predictive modeling, and strategic decision-making in the dynamic landscape of advertising and marketing.

A crucial task of world's biggest search engines, which want to make revenue out of advertising (ads), is to predict impressions of ads, clicks on ads and Click-Through-Rate(CTR) for ads, so that they could show ads to the interested users, according to their search queries. So it is not surprising, that companies like Google and Microsoft invest a lot of money in researches for this field. This paper analyses, how values of impressions, clicks and CTR vary over time. The analysis is done on the open advertisement data set, retrieved from the University College London (UCL). Those three values are also the main focus of this work. We will test, if markets of US and UK are correlated. At the end, we will try to predict CTR value of US-market learned from UK market, using various machine learning techniques The advertising dataset captures the sales revenueenerated with respect to advertisement costs across multiple channels like radio, tv, and newspapers .It is required to understand the impact of ad budgets on the overall sales. Understand the Dataset & clean up (if required).Build Regression models to predict the sales w.r.t a single &. multiple features. Also evaluate the models & compare their respective scores like R2, RMSE, etc.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviations	Full Form
ML	Machine Learning
OLS	Oridary Least Square

CHAPTER1
INTRODUCTION

1.1 OVERVIEW:

In this project, my objective was to analyze sales data. and advertising data, develop a predictive model using linear regression with the OLS method, and draw meaningful conclusions.I began by visually exploring the sales and advertising data, gaining insights into their distributions, patterns, and potential relationships. This step allowed me to identify any outliers, trends, or significant features within the data.To ensure the effectiveness of my model, I performed data scaling and transformation. Scaling techniques, such as normalization or standardization, were applied to bring the variables to a similar scale, preventing any particular feature from dominating the model. Additionally, transformation techniques, such as logarithmic or power transformations, were employed to handle skewed or non-linear relationships.

Using linear regression with the OLS method, I developed a model to predict sales based on the given advertising data. The sales data served as the dependent variable, while the advertising data acted as the predictors. By estimating the coefficients through OLS, I obtained insights into the relationships between advertising expenditures and sales.Throughout the project, I focused on evaluating and validating my model. Key evaluation metrics, including R-squared, adjusted R-squared, and the F-statistic, were utilized to assess the goodness of fit and overall significance of the model. Diagnostic plots, such as residual analysis, were employed to ensure that the assumptions of linear regression were met.Based on my analysis, I concluded that the linearregression model developed using the OLS methodprovided valuable insights into the impact ofadvertising expenditures on sales. Throughappropriate data scaling and transformationtechniques, I enhanced the accuracy and robustnessof the model.

1.2 what is dataset

A dataset is a file that contains one or more records. It is a collection of related, discrete records that are stored and managed as a whole entity. A dataset can be assessed individually or in combination with other team members. It is organized in some type of data structure. Datasets can hold information such as medical records or insurance records. They can be referenced using a name without specifying the location.

A dataset is a collection of raw data collected through research, analyzing the shopping invoice, and sales analysis. The raw data is processed and stored in a repository. Many organizations, such as universities, research agencies, and government institutions, make the datasets on the web for others to download and use for various purposes .Marketing datasets contain consumer data gathered, consolidated, and processed in a systematic manner. These datasets contain data from both the consumers and the potential consumers.

1.3 needs for dataset

Data is a powerful and important marketing tool. It is essential to get the most out of content marketing efforts. Data is the first step in executing your content. Your content is an output of your creativity, and it's trustworthy when your data back it. Content is compelling when you add data, and it is a way of including more backlinks to your content.

For instance, a blog about which social media platform is used frequently by young people is just blank unless it is garnished with statistics about the region, age group, and the type of posts on the social media platform. There are publicly available third-party datasets that you can download for reference and use for the blog article.

1.4 why do we need advertising dataset

The advertising dataset captures the sales revenue generated with respect to advertisement costs across multiple channels like radio, tv, and newspapers. It is required to understand the impact of ad budgets on the overall sales.

Data takes the guesswork out of outdoor and billboard advertising. It helps you optimize your marketing budget, improve your customer experience, and understand which channels, touchpoints, and strategies work.

Market Research:

- They provide insights into consumer behavior, preferences, and trends. This helps businesses understand their target audience better.

Content Creation:

- Understanding what kind of content resonates with the audience can guide the creation of more effective advertisements.

Overall, advertising datasets are a cornerstone of modern marketing, helping businesses reach their target audience effectively and efficiently. However, it's important to handle these datasets ethically, respecting privacy and legal regulations.

1.5 PURPOSE OF THE MACHINE LEARNING

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. Nowadays, with the development of computer technology, pattern recognition has become an essential and important technique in the field of Artificial Intelligence. The pattern recognition can identify letters, images, voice or other objects and also can identify status, extent or other abstractions. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Dataset Description:

The dataset tor this project consists of historical advertising data and comesponding sales figures. It includes information on various advertsing channels, such as TV, radio, and newspaper, along with the corresponding sales numbers recorded during a specific time period.

Data Dictionary:

TV: Advertising expenses in dolars spent on TV advertising.

Radio: Advertising expenses in dolars spart on radio advertising.

Newspaper: Advertsing expenses in dollars spent on newspaper advertising

Sales: Sales figures recorded in units or monetary value comesponding to the advartising efforts.

CHAPTER 2

SYSTEM ANALYSIS

2.1 AIM:

The advertising dataset captures the sales revenue generated with respect to advertisement costs across multiple channels like radio, tv, and newspapers. It is required to understand the impact of ad budgets on the overall sales.

2.2 scope:

1.) Increases awareness:
Advertising's purpose is to make the customer aware of the company and its products. Advertising research on the opposite side is to make the company aware of its target market and target customer, which helps in building effective advertising.

2.) Analyzes changing market:
If a company want to grow in the long run the company needs to know their customer. Customer attitude also changes with the change in the market or environment, because new and innovative products are launched by the companies on a daily bases. Advertising research helps the company to analyzes these changes to know about the changing attitudes of customers.

3.) Advertising Communication
Advertising's purpose is to communicate the product or brand with its target or prospective customer. Successful communication of messages can be measure by increasing awareness about the product, changing the attitude of the customer, taking some action by the customer regarding the product.

4.) Provide Feedback:

Advertising does not end after execution company has to check that they got the desired result or not. It is an attempt to measure that the investment in creating the advertising has resulted in attaining the goals and provide satisfaction to the consumers. Advertising research provides feedback to the company about the effectiveness of advertising.

5.) Provide Results:

Evaluation of advertising refers to the activity of comparing the actual results of advertising to the established standard to know the real value of the advertising performance.It helps to know that message reached the target customers or not. It can be done at any stage, in starting, in the middle, or at the end of the advertising.

2.3 how does it work:

We used libraries such as numpy and pandas for the data science stuff. Matplotlib was used as the graphing library to visualise the data and the seaborn is used for the machine learning operations. We initially classify the features for the model and get the insights of the features visualised. Then make the correlation matrix to check for all the correlations. We then make the data into two sets, one for training and one for testing. Here we can emphasise the importance of the libraries and techniques used. We will talk more about numpy, pandas,linear regression,ordinary least square(OLS),random forest.

Numpy:

Numpy is a fundamental package for scientific computing with Python. It provides a powerful N-dimensional array object and tools for working with these arrays. Numpy is often used in data science to perform mathematical operations on large datasets ,such as computing mean, median, and standard deviation.

Pandas:

Pandas is another important library for data science. It provides data structures for efficiently storing and manipulating large datasets. Pandas allows us to work with tabular data, similar to how we work with data in spreadsheets. We can use Pandas to perform operations such as filtering, sorting, and joining data.

Linear Regression:

Linear regression is a statistical technique used to analyze the relationship between a dependent variable and one or more independent variables. # assumes a linear relationship and finds the best-Sting line to minimize the differences between predicted and actual values. Coefficients represent the strength and direction of influence, while the intercept predicts the value when all variablos are zero. Linear regression is used for prediction, forecasting, and understanding variable impact. Assumptions should be checked before irterpretation and prediction.

Ordinary Least Squares (OLS):

Ordinary Least Squares (OLS) is a widely used method in statistics and econometrics for estimating the parameters of a linear regression model, OLS aims to find the best-fiting line by minimizing the sum of squared residuals between the predicted and actual values. It estimates the regression coefficients that define the relationship between the dependent variable and the independent variables. OLS assumes that the errors are normally distributed with a mean of zero and conart variance. By solving the OLS equations, we ottain cofficient estimates that peovide insights inso the strength and direction of the relationships. OLS is a straighforward and efficient method for linear regression anatyais, making it a fundamental tsol for modeling and inference in various fields.

Random Forest:

Random forest is a popular algorithm for regression and classification problems. It is an ensemble learning method that combines multiple decision trees to make a prediction. Random forest is often used in data science for predicting customer lifetime value, product recommendations, and other regression and classification.

CHAPTER 3

SYSTEM DESIGN AND IMPLEMENTATION

3.1 Steps:

Step 1: Get The Insights Of The Dataset Provided

The first step in any data science project is to gain a thorough understanding of the dataset you are working with. This includes understanding the variables in the dataset, their meaning, and the overall structure of the dataset.

Some key tasks at this stage include:
- Exploring the dataset using summary statistics and visualisations
- Identifying missing values and deciding how to handle them
- Checking for outliers and anomalies in the data
- Assessing the overall quality of the dataset and its suitability for the analysis you wish to perform

Step 2: Select the Features(input variables) For The Model

Once you have a good understanding of the dataset, the next step is to select the features that will be used as inputs for your machine learning model. This involves identifying the variables that are most likely to be predictive of the target variable you are trying to predict.

Some key tasks at this stage include:
- Identifying the target variable that you wish to predict
- Selecting the features that are most relevant to the target variable
- Deciding whether to include all variables in the dataset or only a subset
- Considering the potential interactions between variables and whether to include them in the model

Step 3: Visualise The Data From The Selected Features

Once you have selected your features, it's important to visualise the data to gain a better understanding of the relationships between the variables. This can help you identify patterns in the data and determine whether there are any outliers or anomalies that need to be addressed.

Some key tasks at this stage include:
- Creating scatter plots and histograms to visualise the distribution of each variable
- Creating box plots and violin plots to visualise the relationship between variables
- Creating heat maps and correlation matrices to visualise the relationships between variables and identify potential interactions

Step 4: Preprocessing The Data

Before building the machine learning model, it's important to preprocess the data to ensure that it is in a suitable format for the model. This may involve a range of tasks, depending on the specifics of your dataset and the model you are building.

Some key tasks at this stage include:
- Handling missing values by imputing or removing them
- Scaling the data to ensure that variables are on the same scale
- Encoding categorical variables as humeric values
- Handling outliers and anomalies in the data

Step 5: Build The Machine Learning Model.

With the data preprocessed, it's time to build the machine learning model. There are many different algorithms and techniques available for building machine learning models, so it's important to select the one that is best suited to your dataset and the problem you are trying to solve.

Some key tasks at this stage include:

- Selecting an appropriate algorithm or technique for the model
- Splitting the data into training and testing sets
- Fitting the model to the training data
- Evaluating the performance of the model on the testing data

3.2 SYSTEM REQUIREMENTS

Requirement analysis determines the requirements of a new system. This project analyses product and resource requirements, which is required for this successful system. The product requirement includes input and output requirements it gives the wants in terms of input to produce the required output. The resource requirements give in brief about the software and hardware that are needed to achieve the required functionality.

OS: Windows 7, windows 8, windows, Linux and Mac compatible.

Browser: internet explorer, chrome, firefox and safari (Support all modernbrowsers).

IDE: VS Code (Recommended)

Software: .

Browser: internet explorer, chrome, firefox and safari (Support all modernbrowsers).

Coding language: PYTHON

HARDWARE REQUIREMENT

Processor: Pentium Dual Core 2.00

GHZ**Hard disk:** 120 GB or Above

RAM: 4 GB or Above

Keyboard: 110 keys enhanced

JUPYTER NOTEBOOK:

JupyterLab is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

3.3 ALGORITHMS USED

3.3.1 LINEAR REGRESSION:

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables.

The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

3.3.2 RANDOM FORESTS

Random forest is a popular machine learning algorithm that belongs to the family of ensemble methods. It is widely used for both classification and regression tasks and has gained popularity due to its robustness, scalability, and ease of use.

Random forest combines multiple decision trees to form a forest of trees, where each tree is trained on a randomly selected subset of the data and a randomly selected subset of the features. The random selection of subsets is essential to ensure that the individual trees are diverse and not correlated with each other, which can improve the overall performance of the model.

In conclusion, random forest is a powerful and versatile machine learning algorithm that has gained widespread use in a variety of applications. It combines multiple decision trees to form a robust and scalable model that can handle both classification and regression tasks. Its random selection of subsets helps to reduce overfitting and improve generalisation performance, while its ability to handle missing data and provide feature importance measures makes it suitable for feature selection and interpretation. While random forest has some limitations, its strengths make it an excellent choice for many machine learning problems.
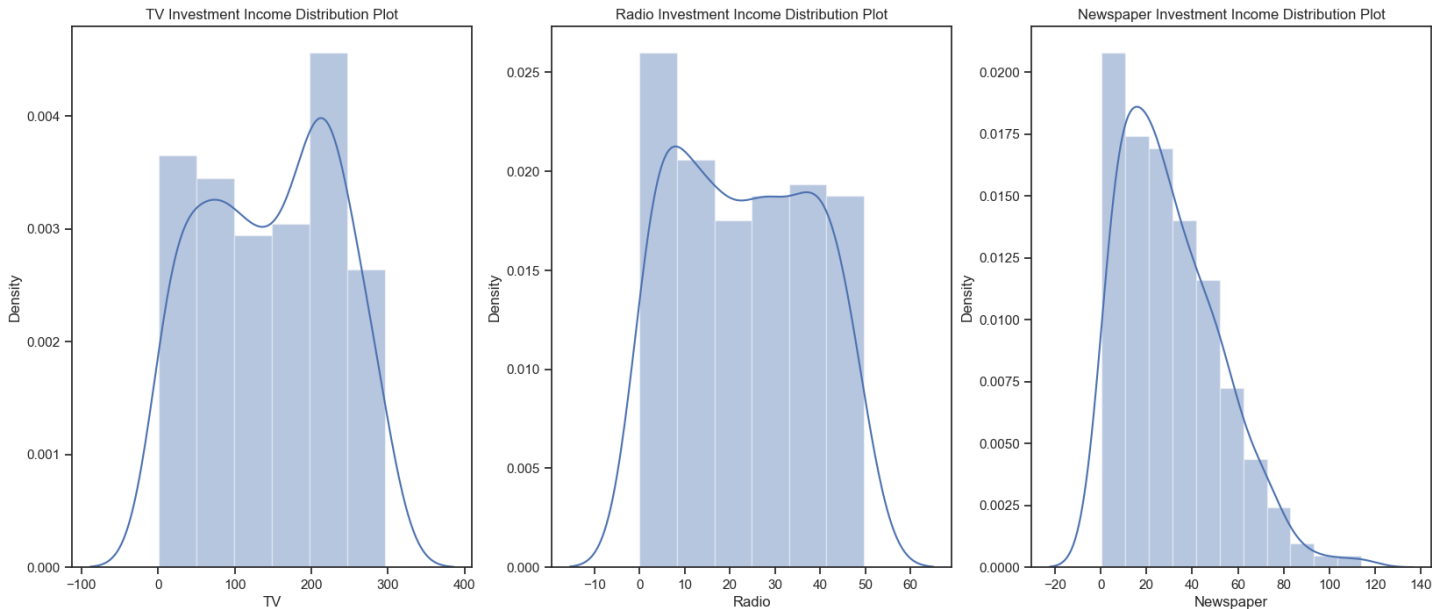
CHAPTER 4

RESULT AND DISCUSSION

FIG 4.1

4.1.1 TV Investment Income Distribution Plot:

the analysis suggests that the linear regression model has a very strong ability to explain the variation in the response variable based on the predictor variables. The high R square value indicates that the model can account for a large proportion of the variation in the response variable. The standard error is relatively low, indicating that the model can make very accurate predictions of the response variable. The adjusted R square value is slightly lower than the R square, which may indicate that the model is slightly overfitting to the data @ that some of the predictor variables may not be significant. None the less, the model appears to be a good fit for the data.

4.1.2 Radio Investment Income Disturbution Plot

the analysis suggests that the linear regression model has a moderate ability to explain the variation in the response variable based on the predictor variables. The moderate R square value indicates that the model can account for a significant proportion of the variation in the response variable. The standard error is relatively low, indicating that the model can make fairly accurate predictions of the response variable. However, the adjusted R square value is slightly lower than the R square, which suggests that the model may be overfitting to the data or that some of the predictor variables may not be significant.

4.1.3 Newspaper Investment Income Distribution Plot:

model has a weak and limited ability to explain the varjation in the response variable based on the predictor variables. The low R square and adjusted R square values indicate that the model does not account for much of the variation in the response variable. The high standard error indicates that the model may not be very accurate in predicting the response variable.
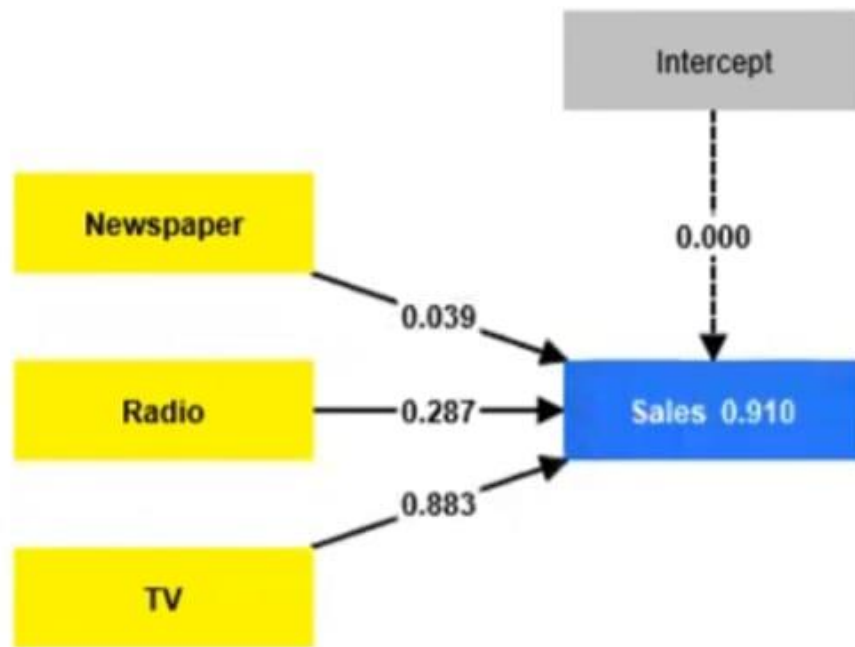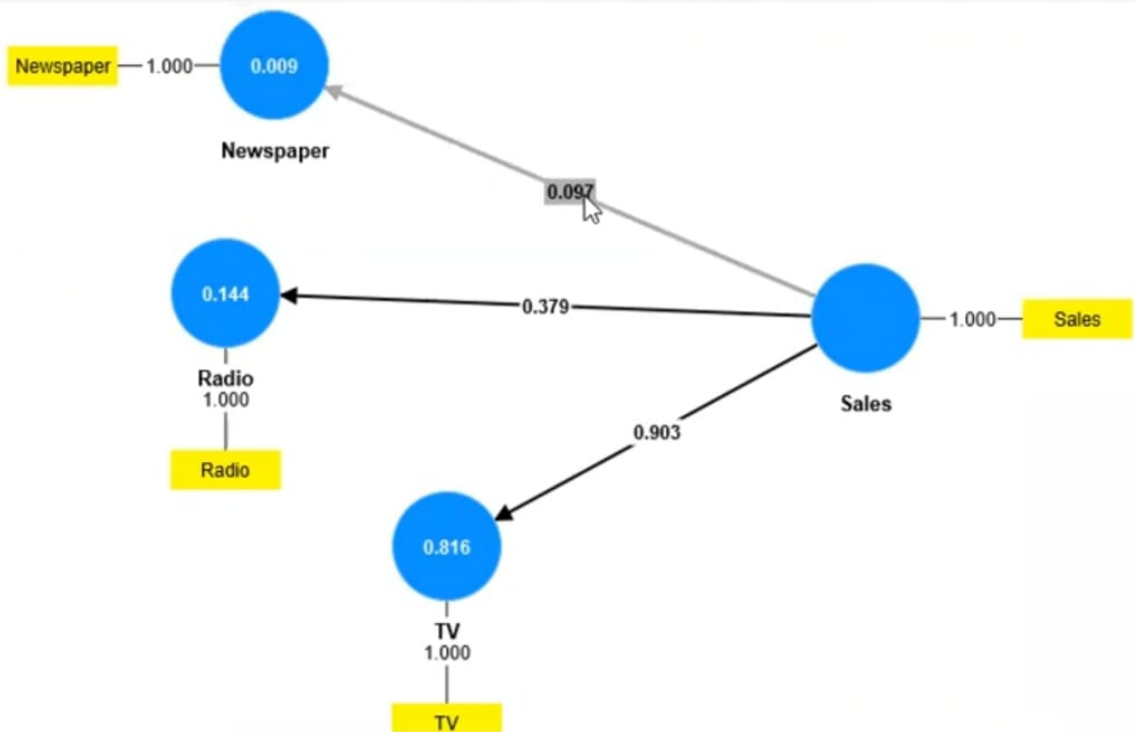

Fig 4.2

Figure 4.3

Figure 4.4

24

CHAPTER 5

CONCLUSION

In this project, our objective was to analyze sales and advertising data, develop a predictive model using linear regression with the OLS method, and draw meaningful conclusions .We began by visually exploring the data, identifying outliers, trends, and significant features. Data scaling techniques like normalization or standardization were applied to ensure an effective model. Additionally, transformation techniques such as logarithmic or power transformations were used to handle skewed or non-linear relationships .Using linear regression with the OLS method, we developed a model to predict sales based on advertising data. The model's coefficients provided insights into the relationships between advertising expenditures and sales. Evaluation metrics like R-squared, adjusted R-squared, and the F-statistic confirmed the model's goodness of fit. Diagnostic plots, including residual analysis, verified the assumptions of linear regressions.

In conclusion our analysis reveled that the linear regression model developed using the OLS method provided variables insights into the impact of the advertising expenditures on sales. Through appropriate data scaling and transformation we improved the models accuracy and robustness

It is important to note that this conclusions on based on specified data and modeling choices limitation and areas for further investing may exist none the less, those findings lay a foundation for understanding the relationship between advertising and sales.

REFERENCE

- D. D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval," in Machine Learning: ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Proceedings, 1998.
- R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1996
- T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," Springer, 2009.
- I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, 2003.
- C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.
- Li, X., Zhang, X., & Li, Q. (2021). Analysis of Employee Turnover Factors Basedon Machine Learning Algorithms: A Case Study of a Large Manufacturing Company. Sustainability, 13(8), 4267
- Amriani, F., Nugroho, L. E., & Siregar, S. M. (2021). Predicting Turnover Intentions Using Supervised Learning Algorithms. Journal of Physics: ConferenceSeries, 1795(1), 012084

SOURCE CODE:

```python
import warnings
warnings.filterwarnings('ignore')

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g.
pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns
# Input data files are available in the read-only "../input/"
directory
# For example, running this (by clicking run or pressing
Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

ad_df = pd.read_csv('advertising.csv')
ad_df.head()

ad_df.shape

ad_df.info()

ad_df.isnull().sum().to_frame().rename(columns={0:'Total No. of
Missing values'})
```

```python
print('Duplicated : ', ad_df.duplicated().sum())

ad_df.describe().T

plt.figure(figsize=(20,8))

plt.subplot(1,3,1)
plt.title('TV Investment Income Distribution Plot')
sns.distplot(ad_df.TV)

plt.subplot(1,3,2)
plt.title('Radio Investment Income Distribution Plot')
sns.distplot(ad_df.Radio)

plt.subplot(1,3,3)
plt.title('Newspaper Investment Income Distribution Plot')
sns.distplot(ad_df.Newspaper)

plt.show()

plt.scatter(ad_df.TV,ad_df.Sales, label='TV')
plt.scatter(ad_df.Radio,ad_df.Sales, label = 'Radio')
plt.scatter(ad_df.Newspaper,ad_df.Sales, label = 'Newspaper')
plt.legend()
plt.title('Invested Income Vs Sales')
plt.xlabel('All Invested Income TV, Radio, Newspaper')
plt.ylabel('Sales')
plt.show()
```

```python
def outliers(column):

    q1 = np.percentile(column,25)
    q3 = np.percentile(column,75)

    IQR = q3-q1

    uF = q3+(1.5*IQR)
    lF = q1-(1.5*IQR)

    mList = list(column)
    list_out = []
    for val in mList:
        if val > uF or val < lF:
            list_out.append(val)

    return list_out,uF,lF

print("Outliers of TV columns : ", outliers(ad_df.TV))
print("Outliers of Radio columns : ",outliers(ad_df.Radio))
print("Outliers of Newspaper columns : ",outliers(ad_df.Newspaper))

listt,upper,lower = outliers(ad_df.Newspaper)

print("Old Shape: ", ad_df.shape)
listt,upper,lower = outliers(ad_df.Newspaper)

upper_array = np.where(ad_df.Newspaper>=upper)[0]
lower_array = np.where(ad_df.Newspaper<=lower)[0]

# Removing the outliers
```

```python
ad_df.drop(index=upper_array, inplace=True)
ad_df.drop(index=lower_array, inplace=True)

print("New Shape: ", ad_df.shape)

sns.set(style="ticks", color_codes=True)
g = sns.pairplot(ad_df)
plt.show()

sns.heatmap(ad_df.corr(),annot=True)
plt.show()

X = ad_df.iloc[:,:-1].values
y = ad_df.iloc[:,-1].values

from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test =
train_test_split(X,y,test_size=0.3,random_state=0)

from sklearn.linear_model import LinearRegression

lr = LinearRegression()
lr.fit(X_train,y_train)

y_pred = lr.predict(X_test)

def accuracy(y_test,y_pred):
sum_rss = 0
   for i in range(len(y_test)):
      diff_rss = (y_test[i]-y_pred[i])**2
      sum_rss = sum_rss+diff_rss
```

```python
    rss  =sum_rss

    y_mean = np.mean(y_test)
    sum_tss = 0
    for i in range(len(y_test)):
        diff_tss = (y_test[i]-y_mean)**2
        sum_tss = sum_tss+diff_tss
    tss = sum_tss

    r2 = 1-(rss/tss)

    return r2

y_test = list(y_test)
y_pred = list(y_pred)
print(accuracy(y_test,y_pred))

from sklearn import metrics
def result_summary(y_true, y_pred):
    explained_variance=metrics.explained_variance_score(y_true, y_pred)
    mean_absolute_error=metrics.mean_absolute_error(y_true, y_pred)

mse=metrics.mean_squared_error(y_true, y_pred)
    mean_squared_log_error=metrics.mean_squared_log_error(y_true,
y_pred)
    median_absolute_error=metrics.median_absolute_error(y_true,
y_pred)
    r2=metrics.r2_score(y_true, y_pred)
    n,p = ad_df.shape
    p=p-1
    adj_r2 = 1-(1-r2)*(n-1)/(n-p-1)
```

```python
    print('explained_variance: ', round(explained_variance,4))
    print('mean_squared_log_error: ',
round(mean_squared_log_error,4))
    print('r2: ', round(r2,4))
    print('adj_r2: ',round(adj_r2,4))
    print('MAE: ', round(mean_absolute_error,4))
    print('MSE: ', round(mse,4))
    print('RMSE: ', round(np.sqrt(mse),4))

result_summary(y_test,y_pred)
```

OUTPUTS:

```
In [21]: ad_df = pd.read_csv('advertising.csv')
         ad_df.head()
```

Out[21]:

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	12.0
3	151.5	41.3	58.5	16.5
4	180.8	10.8	58.4	17.9

```
In [22]: ad_df.shape
```

Out[22]: (200, 4)

```
In [23]: ad_df.info()

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 200 entries, 0 to 199
         Data columns (total 4 columns):
          #    Column       Non-Null Count   Dtype
         ---   ------       --------------   -----
          0    TV            200 non-null     float64
          1    Radio         200 non-null     float64
          2    Newspaper     200 non-null     float64
          3    Sales         200 non-null     float64
         dtypes: float64(4)
         memory usage: 6.4 KB
```

```
In [24]: ad_df.isnull().sum().to_frame().rename(columns={0:'Total No. of Missing values'})
Out[24]:
```

|	Total No. of Missing values
TV	0
Radio	0
Newspaper	0
Sales	0

```
In [25]: print('Duplicated : ', ad_df.duplicated().sum())

         Duplicated :  0
```
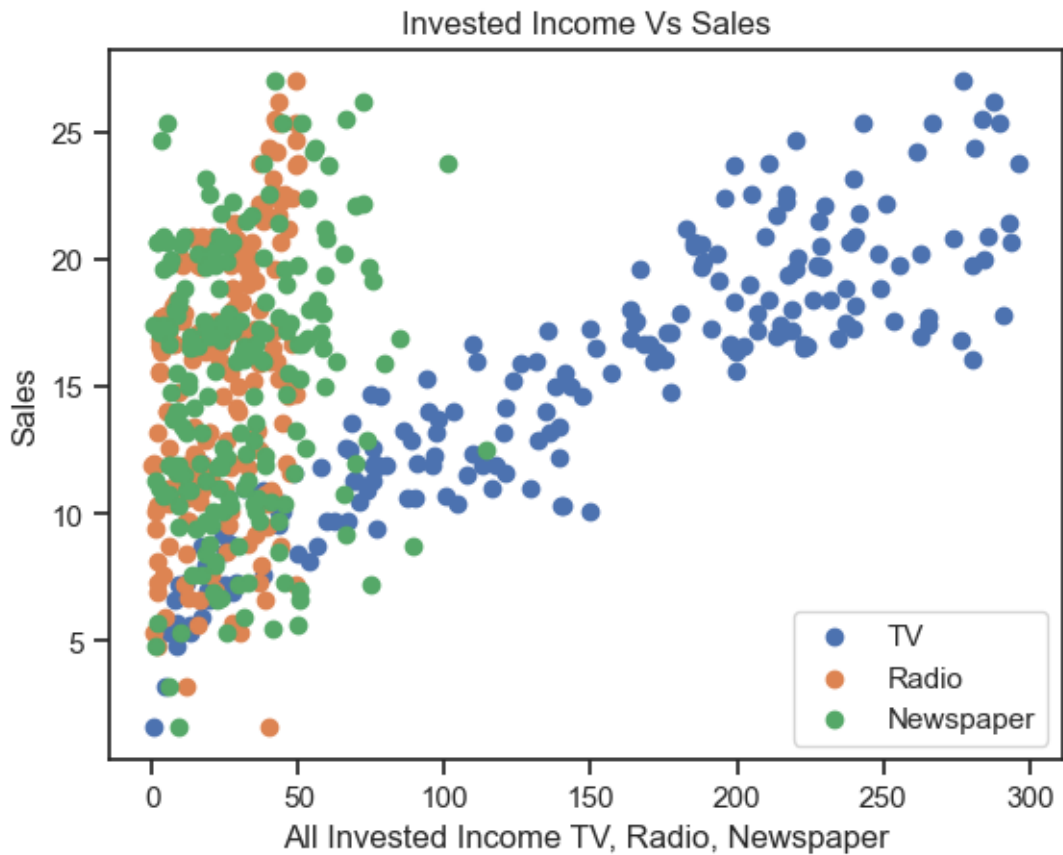
```
In [26]: ad_df.describe().T
```

Out[26]:

|	count	mean	std	min	25%	50%	75%	max
TV	200.0	147.0425	85.854236	0.7	74.375	149.75	218.825	296.4
Radio	200.0	23.2640	14.846809	0.0	9.975	22.90	36.525	49.6
Newspaper	200.0	30.5540	21.778621	0.3	12.750	25.75	45.100	114.0
Sales	200.0	15.1305	5.283892	1.6	11.000	16.00	19.050	27.0

Invested Income Vs Sales
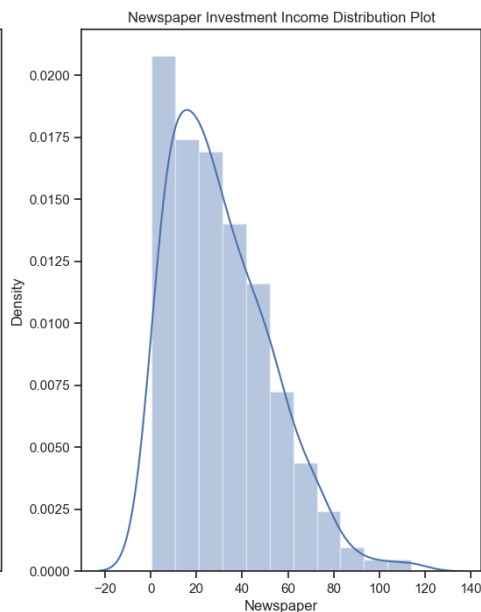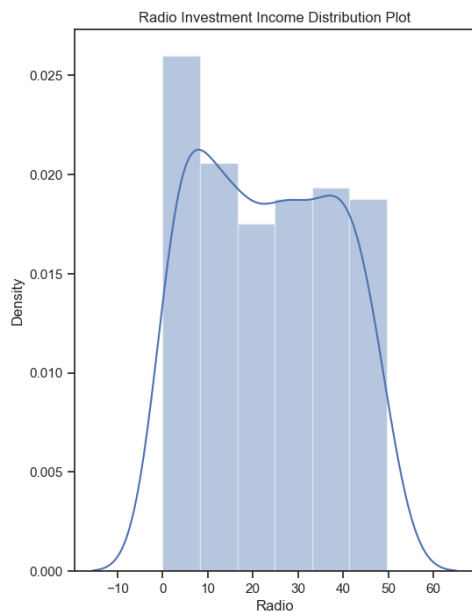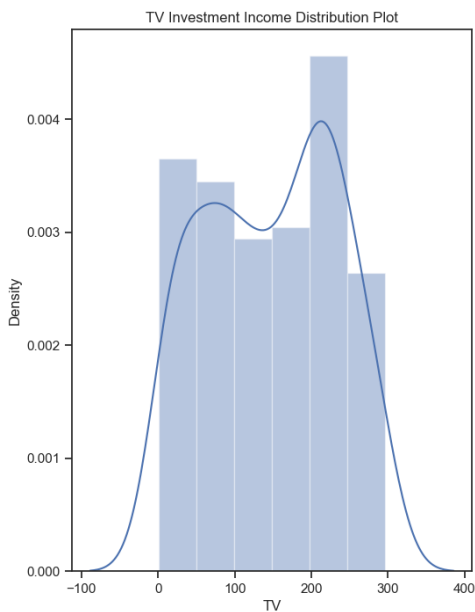
Outliers of TV columns : ([], 435.5, -142.29999999999998)

Outliers of Radio columns: ([], 76.35, 29.849999999999994)

Outliers of Newspaper columns: ([114.0, 100.91, 93.625, -35.775000000000006)
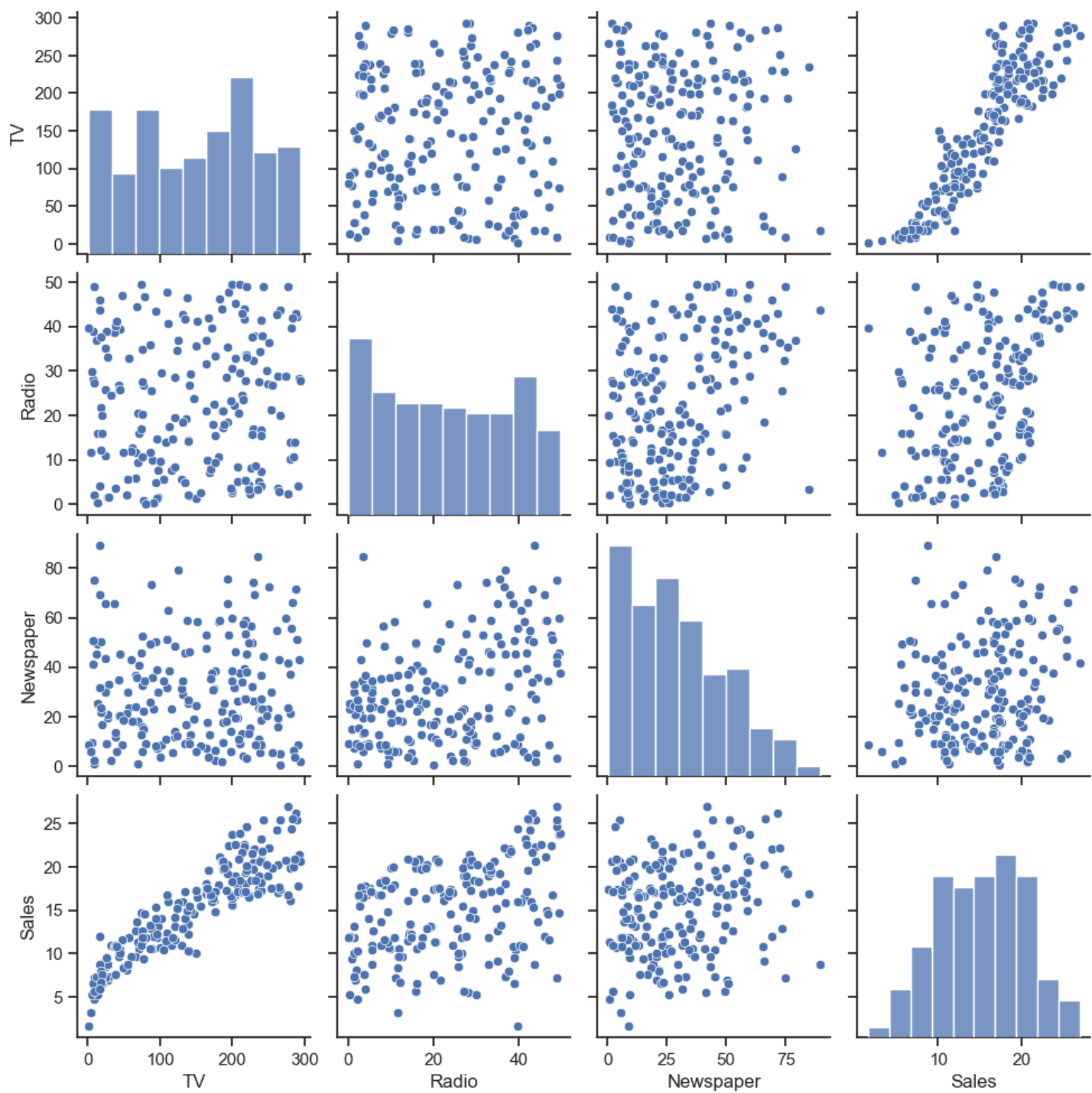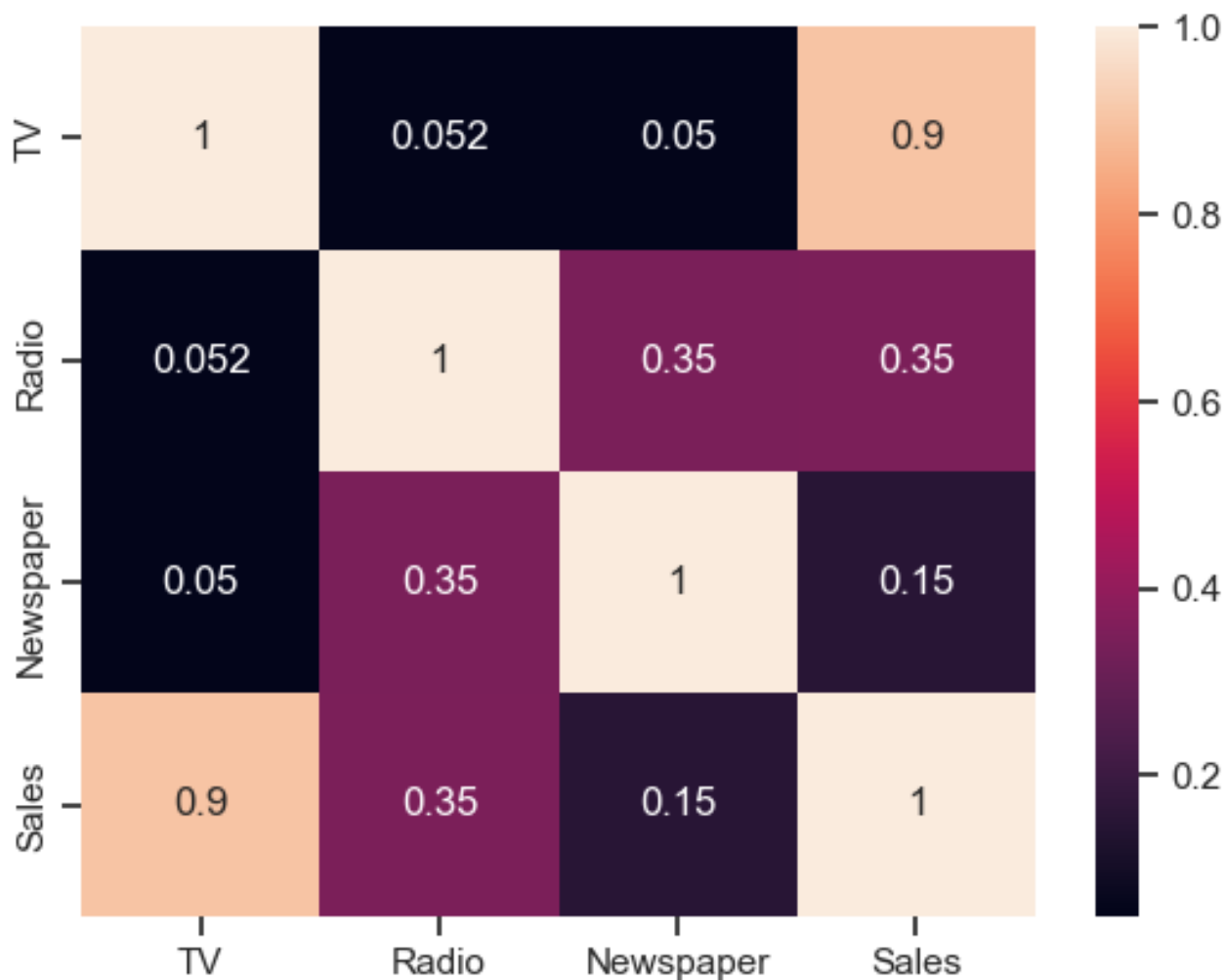
```
In [36]:  from sklearn.linear_model import LinearRegression

          lr = LinearRegression()
          lr.fit(X_train,y_train)

Out[36]:  ▼ LinearRegression
          LinearRegression()
```

explained_variance: 0.9206
mean_ squared _log_error: 0.0122
r2: 0.9189
adj_r2: 0.9177
MAE: 1.1891
MSE: 2.5245
RMSE: 1.5889