Chapter 1
Learning Objective of Internship

- To learn what is Data Science and Business Analytics.

- To learn about python and its libraries like-

1. Pandas
2. Matplotlib
3. Numpy
4. Sk-learn
5. seaborn

- To learn about Exploratory Data Analysis.

- To design some visuals and perform Data Visualization.

- To learn Simple Linear Regression.

- To make some Machine Learning Model

Chapter 2
WEEKLY OVERVIEW OF INTERNSHIP ACTIVITIES

	DATE	DAY	Name of the Topic/Module Completed
	28-07-2022	Thursday	Orientation
	29-07-2022	friday	Introduction of internship

1st Week	DATE	DAY	Name of the Topic/Module Completed
	02-07-2022	Tuesday	Data science
	03-07-2022	Wednesday	Business analytics
	04-07-2022	Thursday	Machine learning
	05-07-2022	Friday	Simple Linear Regression
	06-07-2022	Saturday	Python

2nd Week	DATE	DAY	Name of the Topic/Module Completed
	08-07-2022	Monday	Sk-learn
	09-07-2022	Tuesday	Pandas
	10-07-2022	Wednesday	Matplotlib
	11-07-2022	Thursday	Seaborn
	12-07-2022	Friday	Numpy

	13-07-2022	**Saturday**	**Practice assessment**

3rd Week	**DATE**	**DAY**	**Name of the Topic/Module Completed**
	15-07-2022	**Monday**	**Project Introduction**
	16-07-2022	**Tuesday**	**Project Making**

4th Week	**DATE**	**DAY**	**Name of the Topic/Module Completed**
	18-07-2022 - 23-07-2022		**Task 1**
	24-07-2022- 29-07-2022		**Task 2**

5th Week	**DATE**	**DAY**	**Name of the Topic/Module Completed**
	30-07-2022	**Tuesday**	**Project submission**

Chapter 3
INTRODUCTION

Data Science-

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning.

The accelerating volume of data sources, and subsequently data, has made data science is one of the fastest growing field across every industry. As a result, it is no surprise that the role of the data scientist was dubbed the "sexiest job of the 21st century" by Harvard Business Review (link resides outside of IBM). Organizations are increasingly reliant on them to interpret data and provide actionable recommendations to improve business outcomes.

The data science lifecycle involves various roles, tools, and processes, which enables analysts to glean actionable insights. Typically, a data science project undergoes the following stages:

Data ingestion: The lifecycle begins with the data collection--both raw structured and unstructured data from all relevant sources using a variety of methods. These methods can include manual entry, web scraping, and real-time streaming data from systems and devices. Data sources can include structured data, such as customer data, along with unstructured data like log files, video, audio, pictures, the Internet of Things (IoT), social media, and more.

Data storage and data processing: Since data can have different formats and structures, companies need to consider different storage systems based on the type of data that needs to be captured. Data management teams help to set standards around data storage and structure, which facilitate workflows around analytics, machine learning and deep learning models. This stage includes cleaning data, deduplicating, transforming and combining the data using ETL (extract, transform, load) jobs or other data integration technologies. This data preparation is essential for promoting data quality before loading into a data warehouse, data lake, or other repository.

Data analysis: Here, data scientists conduct an exploratory data analysis to examine biases, patterns, ranges, and distributions of values within the data. This data analytics exploration drives hypothesis generation for a/b testing. It also allows analysts to determine the data's relevance for use within modeling efforts for predictive analytics, machine learning, and/or deep learning. Depending on a model's accuracy, organizations can become reliant on these insights for business decision making, allowing them to drive more scalability.

Communicate: Finally, insights are presented as reports and other data visualizations that make the insights—and their impact on business—easier for business analysts and other decision-makers to understand. A data science programming language such as R or Python includes components for generating visualizations; alternately, data scientists can use dedicated visualization tools.

Data scientists rely on popular programming languages to conduct exploratory data analysis and statistical regression. These open source tools support pre-built statistical modeling, machine learning, and graphics capabilities. These languages include the following (read more at "Python vs. R: What's the Difference?"):

R Studio: An open source programming language and environment for developing statistical computing and graphics.
Python: It is a dynamic and flexible programming language. The Python includes numerous libraries, such as NumPy, Pandas, Matplotlib, for analyzing data quickly.
To facilitate sharing code and other information, data scientists may use GitHub and Jupyter notebooks.

Some data scientists may prefer a user interface, and two common enterprise tools for statistical analysis include:

SAS: A comprehensive tool suite, including visualizations and interactive dashboards, for analyzing, reporting, data mining, and predictive modeling.

IBM SPSS: Offers advanced statistical analysis, a large library of machine learning algorithms, text analysis, open source extensibility, integration with big data, and seamless deployment into applications.

Data scientists also gain proficiency in using big data processing platforms, such as Apache Spark, the open source framework Apache Hadoop, and NoSQL databases. They are also skilled with a wide range of data visualization tools, including simple graphics tools included with business presentation and spreadsheet applications (like Microsoft Excel), built-for-purpose commercial visualization tools like Tableau and IBM Cognos, and open source tools like D3.js (a JavaScript library for creating interactive data visualizations) and RAW Graphs. For building machine learning models, data scientists frequently turn to several frameworks like PyTorch, TensorFlow, MXNet, and Spark MLib.

Given the steep learning curve in data science, many companies are seeking to accelerate their return on investment for AI projects; they often struggle to hire the talent needed to realize data science project’s full potential. To address this gap, they are turning to multipersona data science and machine learning (DSML) platforms, giving rise to the role of “citizen data scientist.”

Multipersona DSML platforms use automation, self-service portals, and low-code/no-code user interfaces so that people with little or no background in digital technology or expert data science can create business value using data science and machine learning. These platforms also support expert data scientists by also offering a more technical interface. Using a multipersona DSML platform encourages collaboration across the enterprise.
Enterprises can unlock numerous benefits from data science. Common use cases include process optimization through intelligent automation and enhanced targeting and personalization to improve the customer experience (CX). However, more specific examples include:

Here are a few representative use cases for data science and artificial intelligence:

An international bank delivers faster loan services with a mobile app using machine learning-powered credit risk models and a hybrid cloud computing architecture that is both powerful and secure.

An electronics firm is developing ultra-powerful 3D-printed sensors to guide tomorrow's driverless vehicles. The solution relies on data science and analytics tools to enhance its real-time object detection capabilities.

A robotic process automation (RPA) solution provider developed a cognitive business process mining solution that reduces incident handling times between 15% and 95% for its client companies. The solution is trained to understand the content and sentiment of customer emails, directing service teams to prioritize those that are most relevant and urgent.

A digital media technology company created an audience analytics platform that enables its clients to see what's engaging TV audiences as they're offered a growing range of digital channels. The solution employs deep analytics and machine learning to gather real-time insights into viewer behavior.

An urban police department created statistical incident analysis tools to help officers understand when and where to deploy resources in order to prevent crime. The data-driven solution creates reports and dashboards to augment situational awareness for field officers.
Shanghai Changjiang Science and Technology Development used IBM® Watson® technology to build an AI-based medical assessment platform that can analyze existing medical records to categorize patients based on their risk of experiencing a stroke and that can predict the success rate of different treatment plans.

Business analytics-

Business analytics, a data management solution and business intelligence subset, refers to the use of methodologies such as data mining, predictive analytics, and statistical analysis in order to analyze and transform data into useful information, identify and anticipate trends and outcomes, and ultimately make smarter, data-driven business decisions.

The main components of a typical business analytics dashboard include:

Data Aggregation: prior to analysis, data must first be gathered, organized, and filtered, either through volunteered data or transactional records
Data Mining: data mining for business analytics sorts through large datasets using databases, statistics, and machine learning to identify trends and establish relationships
Association and Sequence Identification: the identification of predictable actions that are performed in association with other actions or sequentially
Text Mining: explores and organizes large, unstructured text datasets for the purpose of qualitative and quantitative analysis
Forecasting: analyzes historical data from a specific period in order to make informed estimates that are predictive in determining future events or behaviors
Predictive Analytics: predictive business analytics uses a variety of statistical techniques to create predictive models, which extract information from datasets, identify patterns, and provide a predictive score for an array of organizational outcomes
Optimization: once trends have been identified and predictions have been made, businesses can engage simulation techniques to test out best-case scenarios
Data Visualization: provides visual representations such as charts and graphs for easy and quick data analysis

Using business analytics tools
Business data analytics has many individual components that work together to provide insights. While business analytics tools handle the elements of crunching data and creating insights through reports and visualization, the process actually starts with the infrastructure for bringing that data in. A standard workflow for the business analytics process is as follows:

Data collection: Wherever data comes from, be it IoT devices, apps, spreadsheets, or social media, all of that data needs to get pooled and centralized for access. Using a cloud database makes the collection process significantly easier.

Data mining: Once data arrives and is stored (usually in a data lake), it must be sorted and processed. Machine learning algorithms can accelerate this by recognizing patterns and repeatable actions, such as establishing metadata for data from specific sources, allowing data scientists to focus more on deriving insights rather than manual logistical tasks.

Descriptive analytics: What is happening and why is it happening? Descriptive data analytics answers these questions to build a greater understanding of the story behind the data.

Predictive analytics: With enough data—and enough processing of descriptive analytics—business analytics tools can start to build predictive models based on trends and historical context. These models can thus be used to inform future decisions regarding business and organizational choices.

Visualization and reporting: Visualization and reporting tools can help break down the numbers and models so that the human eye can easily grasp what is being presented. Not only does this make presentations easier, these types of tools can help anyone from experienced data scientists to business users quickly uncover new insights.

Business analytics vs. business intelligence

On the face of it, there may not seem to be much difference between business analytics and business intelligence. Some overlap does exist between the two, but looking at business analytics versus business intelligence still creates a gap that needs some explanation.

Certainly, the terms are extremely connected, but business intelligence uses historical and current data to understand what happened in the past and what is happening now. Business analytics, on the other hand, builds on the foundation of business intelligence and attempts to make educated predictions about what might happen in the future. In order to make data-driven predictions about the likelihood of future outcomes, business analytics uses next-generation technology, such as machine learning, data visualization, and natural language query.

Benefits of business analytics

Business analytics benefits impact every corner of your organization. When data across departments consolidates into a single source, it syncs up everyone in the end-to-end process. This ensures there are no gaps in data or communication, thus unlocking benefits such as:

Data-driven decisions: With business analytics, hard decisions become smarter—and by smart, that means that they are backed up by data. Quantifying root causes and clearly identifying trends creates a smarter way to look at the future of an organization, whether it be HR budgets, marketing campaigns, manufacturing and supply chain needs, or sales outreach programs.

Easy visualization: Business analytics software can take unwieldy amounts of data and turn it into simple-yet-effective visualizations. This accomplishes two things. First, it makes insights much more accessible for business users with just a few clicks. Second, by putting data in a visual format, new ideas can be uncovered simply by viewing the data in a different format.

Modeling the what-if scenario: Predictive analytics creates models for users to look for trends and patterns that will affect future outcomes. This previously was the domain of experienced data scientists, but with business analytics software powered by machine learning, these models can be generated within the platform. That gives business users the ability to quickly tweak the model by creating what-if scenarios with slightly different variables without any need to create sophisticated algorithms.

Go augmented: All of the points above consider the ways that business data analytics expedite user-driven insights. But when business analytics software is powered by machine learning and artificial intelligence, the power of augmented analytics is unlocked. Augmented analytics uses the ability to self-learn, adapt, and process bulk quantities of data to automate processes and generate insights without human bias.

Business analytics use cases

More and more departments are trying to better understand how their decisions and budgets affect the business at large. With business analytics software, it's possible to use data to drive strategic decisions, regardless of task or department:

Marketing: Analytics to identify success and impact

Which customers are more likely to respond to an email campaign? What was the last campaign's ROI? More and more marketing departments are trying to better understand how their programs affect the business at large. With AI and machine learning powering analysis, it's possible to use data to drive strategic marketing decisions. Learn more

Human Resources: Analytics to find and share talent insights

What actually drives employee decisions regarding their career? More and more HR leaders are trying to better understand how their programs affect the business at large. With the right analytical capabilities, HR leaders are able to quantify and predict outcomes, understand recruitment channels, and review employee decisions en masse. Learn more

Sales: Analytics to optimize your sales

What is the critical moment that converts a lead to a sale? In-depth analytics can break down the sales cycle, taking in all of the different variables that lead to a purchase. Price, availability, geography, season, and other factors can be the turning point on the customer journey—and analytics offer the tool to decipher that key moment. Learn more

Finance: Analytics to power predictive organizational budgets

How can you increase your profit margins? Finance works with every department, be it HR or sales. That means that innovation is always key, especially as finance departments face larger volumes of data. With analytics, it's possible to bring finance into the future for predictive modeling, detailed analysis, and insights from machine learning.

Exploratory Data Analysis-

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

In statistics, exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing. Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA),which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

Specific statistical functions and techniques you can perform with EDA tools include:

- Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables.
- Univariate visualization of each field in the raw dataset, with summary statistics.
- Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you're looking at.
- Multivariate visualizations, for mapping and understanding interactions between different fields in the data.
- K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e. the number of clusters, based on the distance from each group's centroid.
- The data points closest to a particular centroid will be clustered under the same category.
- K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression.
- Predictive models, such as linear regression, use statistics and data to predict outcomes.

Simple Linear Regression-

Simple Linear Regression is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable. The relationship shown by a Simple Linear Regression model is linear or a sloped straight line, hence it is called Simple Linear Regression.

The key point in Simple Linear Regression is that the dependent variable must be a continuous/real value. However, the independent variable can be measured on continuous or categorical values.

Simple Linear regression algorithm has mainly two objectives:

Model the relationship between the two variables. Such as the relationship between Income and expenditure, experience and Salary, etc.
Forecasting new observations. Such as Weather forecasting according to temperature, Revenue of a company according to the investments in a year, etc.

When to use regression

We are often interested in understanding the relationship among several variables. Scatterplots and scatterplot matrices can be used to explore potential relationships between pairs of variables. Correlation provides a measure of the linear association between pairs of variables, but it doesn't tell us about more complex relationships. For example, if the relationship is curvilinear, the correlation might be near zero.

You can use regression to develop a more formal understanding of relationships between variables. In regression, and in statistical modeling in general, we want to model the relationship between an output variable, or a response, and one or more input variables, or factors.

Depending on the context, output variables might also be referred to as dependent variables, outcomes, or simply Y variables, and input variables might be referred to as explanatory variables, effects, predictors or X variables.

We can use regression, and the results of regression modeling, to determine which variables have an effect on the response or help explain the response. This is known as explanatory modeling.

We can also use regression to predict the values of a response variable based on the values of the important predictors. This is generally referred to as predictive modeling. Or, we can use regression models for optimization, to determine settings of factors to optimize a response. Our optimization goal might be to find settings that lead to a maximum response or to a minimum response. Or the goal might be to hit a target within an acceptable window.

For example, let's say we're trying to improve process yield.

We might use regression to determine which variables contribute to high yields,
We might be interested in predicting process yield for future production, given values of our predictors, or
We might want to identify factor settings that lead to optimal yields.
We might also use the knowledge gained through regression modeling to design an experiment that will refine our process knowledge and drive further improvement.

Python-

Python is a very popular general-purpose interpreted, interactive, object-oriented, and high-level programming language. Python is dynamically-typed and garbage-collected programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL).

Python supports multiple programming paradigms, including Procedural, Object Oriented and Functional programming language. Python design philosophy emphasizes code readability with the use of significant indentation.

Python is consistently rated as one of the world's most popular programming languages. Python is fairly easy to learn, so if you are starting to learn any programming language then Python could be your great choice. Today various Schools, Colleges and Universities are teaching Python as their primary programming language. There are many other good reasons which makes Python as the top choice of any programmer:

- Python is Open Source which means its available free of cost.
- Python is simple and so easy to learn
- Python is versatile and can be used to create many different things.
- Python has powerful development libraries include AI, ML etc.
- Python is much in demand and ensures high salary
-

Python is a MUST for students and working professionals to become a great Software Engineer specially when they are working in Web Development Domain. I will list down some of the key advantages of learning Python:

- Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- Python is Interactive − You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- Python is Object-Oriented − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
- Python is a Beginner's Language − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

Chapter 4
Details of Internship

TRAINING-

Two weeks of Python training is done learned some libraries -

- Pandas
- Matplotlib
- Numpy
- Sk-learn
- Seaborn

TASKS-
Two task was given in which I have to perform -

1-Exploratory Data Analysis on retail

In this EDA task, I performed 'Exploratory Data Analysis' on 'SampleSuperstore'dataset . As a business analyst , I tried to find out the weak areas where i can work to make it more profitable . Also , what all business problem can be derived by exploring the data.

Questions to be solved-

- What products do the most profit making states buy?
- What products do the loss bearing states buy?
- What products segment needs to be improved in order to drive the profit higher

Work done on this project-

- IMPORTING LIBRARIES
- READING THE DATASET
- DATA PREPROCESSING

- DATA ANALYSIS and VISUALIZATION

2- Prediction using Supervised Learning

SIMPLE LINEAR REGRESSION

In this regression task I predicted the percentage of marks that a student is expected to score based upon the number of hours they studied. This is a simple linear regression task as it involves just two variables.

Questions to be solved-

- What will be the predicted score of a student if he/she studies for 9.25 hrs/day?
- Check the R squared value to check the accuracy of our mode

Work done on this project-

- IMPORTING LIBRARIES
- READING THE DATASET
- DATA PREPROCESSING
- PREPARING THE DATA
- TRAINING THE ALGORITHM
- PREDICTING DATASET
- SOLVING THE TASK USING MODEL
- EVALUATING THE MODEL

Chapter 5
Code of the Internship Taks

Code of Exploratory Data Analysis on retail

Importing Libraries

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Reading the dataset

```
# read dataset 'SampleSuperstore'
sample=
pd.read_csv('C:\\Users\\kreez\\Desktop\\GRIPInsternship\\First_project\\SampleSuperstore.csv')

#Checking shape of dataset
sample.shape
```

Data preprocessing

```
#top 5 rows
sample.head()

##Dataset Description
#Pandas describe():
##is used to view some basic statistical details like percentile,mean,std etc.of a data frame or a series of numeric value.

sample.describe()


#Counting category

sample['Category'].value_counts()
```

```
#Finding basic information

sample.info()

# check missing values

sample.isnull().sum()

# checking for the duplicate data - if any then dropping those data

sample.duplicated().sum()


#droping duplicate data

sample.drop_duplicates()

# display unique data

sample.nunique()

#drop irrelevant columns

col=['Postal Code']
sampleSS = sample.drop(columns= col,axis = 1)

# correlation between variables

sampleSS.corr()

# covariance of columns

sampleSS.cov()

# load first 5 rows of sampleSS

sampleSS.head()
```

Data Analysis and Visualization

```
# visualizing Category and sub category using bar graph

plt.figure(figsize=(16,8))

plt.bar("Sub-Category","Category",data=sampleSS, color= "pink")
plt.title("Category Vs Sub-Category",fontsize=20)
plt.xlabel('Sub-category',font size=15)
plt.ylabel("Category",font size=15)
plt.xticks(rotation=45)
plt.show()


# Histogram for discount, profit, quality and sales

sampleSS.hist(bins=50,figsize=(20,15))
plt.show()

# count the total repeatable states

sampleSS['State'].value_counts()


# now, ploting this state data in countplot

plt.figure(figsize=(16,16))

sns.countplot(x=sampleSS['State'])
plt.title("STATE")
plt.xticks(rotation=90)

# Analysis of Sub-Category

plt.figure(figsize=(12,10))
sampleSS['Sub-Category'].value_counts().plot.pie(autopct="%1.1f%%")
plt.show()plt.show()



# Analysis for total Profit and sales based on Sub-Category

sampleSS.groupby('Sub-Category')['Profit','Sales'].agg(['sum']).plot.bar()
plt.title('Total Profit and Sales per sub-category')
plt.show()
```

```
sns.set(style="whitegrid")
plt.figure(2,figsize=(16,8))

sns.barplot(x='Sub-Category',y='Profit',data=sample, palette='Spectral')
plt.suptitle('Pie consumption pattern in the united state',fontsize=20)

plt.show()


# Analysis based on Shipping mode

sns.countplot(x=sample['Ship Mode'])



# Plotting pair plot for Sub-Category

figsize=(15,10)
sns.pairplot(sampleSS,hue='Sub-Category')
plt.show()



# Now, ploting line plot for discount and profit

plt.figure(figsize=(10,4))
sns.lineplot('Discount','Profit',data=sampleSS,color='y', label='Discount')
plt.legend()
plt.show()


#Plots the turnover generated by different product Categories and Sub-categoies for the list of given states.

def state_data_viewer(state):
    product_data= sampleSS.groupby(['State'])
    for state in states:
        data= product_data.get_group(state).groupby(['Category'])
        fig,ax=plt.subplots(1,3,figsize=(28,5))
        fig.suptitle(state,fontsize=14)
        ax_index=0
        for cat in ['Furniture','Office Supplies','Technology']:
            cat_data=data.get_group(cat).groupby(['Sub-Category']).sum()
            sns.barplot(x=cat_data.Profit,y=cat_data.index,ax=ax[ ax_index])
            ax[ ax_index].set_ylabel(cat)
            ax_index +=1
```

```
    fig.show()

states=['California','Washington','Mississippi','Arizona','Texas']
state_data_viewer(states)
```

Code of Prediction using Supervised Learning-

Importing Libraries

```
# Importing all libraries required
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error
%matplotlib inline
```

Reading the dataset

```
# Reading the given dataset
url = "http://bit.ly/w-data"
data = pd.read_csv(url)
print("Data imported successfully")

data.head()
```

Data preprocessing

```
# using dataframe.shape to get a tuple representing the dimensionality of the DataFrame (Rows,Columns)
print (data.shape)

# using dataframe.describe to get the statistical summary of the DataFrame
print (data.describe())


# using dataframe.info to get a concise summary of the DataFrame
print (data.info())
```

```
# checking the presence of missing values in the datafarme
data.isnull().values.any()


#Let's visualize the dataset by plotting a 2-D graph to check if there is any relationship between the data.
# visualizing the dataframe using a scatterplot
data.plot(x='Hours', y='Scores', style='o')
plt.title('Hours vs Percentage')
plt.xlabel('Hours Studied')
plt.ylabel('Percentage Score')
plt.grid (True)
plt.show()
```

Preparing the data

```
# dividing the dataset into dependent(input) and independent(output) variables
x = data['Hours']     #Input
y = data['Scores']    #Output
print ("(Index)(Value of x) \n")
print (x)
print ("(Index)(Value of y) \n")
print (y)
```

Training the algorithm

```
X = data.iloc[:, :-1].values
Y = data.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=0.2, random_state=0)
lm =LinearRegression()
model=lm.fit(X_train, Y_train)
print("Intercept:                    ", model.intercept_)
print("Coefficient of the independent variable: ",model.coef_)
print("\nTraining complete.")


from sklearn.linear_model import LinearRegression
LR = LinearRegression()
LR.fit(X_train, Y_train)
```

```
print("Data is Trained")
Data is Trained
#Regression Line
line = model.intercept_ + model.coef_*x
line = LR.coef_ * X+LR.intercept_

plt.scatter(X,Y,color='blue', label = 'Score')
plt.title("Study Hours Vs Percentage scores \n Graph")
plt.xlabel("Hours")
plt.ylabel("Scores")
plt.plot(X,line,color='green', label = 'regression line')
plt.legend()
plt.show()
```

PREDICTING DATASET

```
Y_predicted = model.predict(X_test)

# Comparing Actual vs Predicted
df = pd.DataFrame({'Actual': Y_test, 'Predicted': Y_predicted})
df
```

SOLVING THE TASK USING MODEL

```
task = 9.25 #hrs/day
predicted_score = model.predict([[task]])
print("No. of Hours Studied :      ", task)
print("Predicted Percentage Score :",predicted_score[0])



from sklearn.metrics import r2_score
Y_true = [20,27,69,30,62]
Y_predicted = [16.9,33.8,75.4,26.8,60.5]
r2_score(Y_true,Y_predicted)
```

Evaluating the model

```
# Evaluating absolute errors
from sklearn import metrics
print("Mean Absolute Error :", metrics.mean_absolute_error(Y_test, Y_predicted))
print("Mean Squared Error :", metrics.mean_squared_error(Y_test, Y_predict)
```

Chapter 6
Output of Internship Tasks

Output of the Exploratory Data Analysis on retail -

Analysis result

The graph that can visualize that "Blinders" Sub-Category has suffered the highest amount of loss and also profit amongst all other Sub-Categories(for now we can say that what reason for this but may be because of discount given on blinders Sub-Category).

"Copies" has gained highest amount of profit with no loss. There are other Subcategories too haven't faced any kind of losses but their profit margins are also low.

So,Machines suffering highest loss.

The plot we can say that our data is not normal and it has outlier.

The Data Visualization, I see the states and the category where sales and profit are high or less, we can improve in the those states by providing discounts in prefered range so that the company and consumer with both be in profit.

Here, while the superstore is incurring losses by providing discount on their products, they can't stop doing so. Most of the heavy discount are during festival, end-of-season and clearance sales which are necessary, so that the store can make space in their warehouses for fresh stock.

Also, by incurring small losses, the company gains in the future by attracting more long term customers. Therefore, the small losses from discounts are an essential part of company's businesses.

Output of the Prediction using Supervised Learning-

Model predicts that if a student studies for 9.25 hrs/day then he/she will score 93.69%.

The R Squared value is 0.944. It can be referred that 94.4% of the changeability of the dependent output attribute can be explained by the model which means it can give 94.4% accurate results!

The accuracy of our model is 94.4%

Chapter 7
Conclusion

In a nutshell, this internship has been an excellent and rewarding experience. I can conclude that there have been a lot I've learnt from my work at The Spark Foundation.

Needless to say, the technical aspects of the work I've done are not flawless and could be improved provided enough time.
As someone with no prior experience with Python whatsoever I believe my time spent in research and discovering it was well worth it and contributed to finding an acceptable solution to build a fully Data analysis.

Two main things that I've learned the importance of are time-management skills and self-motivation.
I have learned how to put my knowledge and skills into practice, the benefits of networking, understanding the workplace culture and many more.

As an intern, good communication will help with productivity, efficiency, engagement, and growth.

I am grateful to have had the opportunity to do an training and have received so much value from what I learned.

Chapter 8

References

Github-
https://github.com/Devendra-pawar/GRIP_Intern_Projects.git

The Spark Foundation-
https://www.thesparksfoundationsingapore.org/

Google-
https://www.Google.co.in

GFG-
https://www.geeksforgeeks.org/