

Summer Training
Machine Learning & Data Science Program
at IBM

A training report

Submitted in partial fulfillment of the requirements for the award of degree of

Computer Science and Engineering
(Artificial Intelligence and Machine Learning)

Submitted to

LOVELY PROFESSIONAL UNIVERSITY
PHAGWARA, PUNJAB



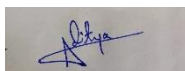
From 06/07/23 to 07/19/23

SUBMITTED BY

Name of the student: Aditya Sajja

Registration Number:12107787

Signature of the student:



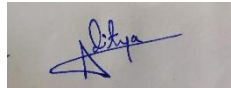
Student Declaration

To whom so ever it may concern

I, **Aditya Sajja,12107787** hereby declare that the work done by me on “**Data Science**” from **June,2023** to **July,2023**, is a record of original work for the partial fulfillment of the requirements for the award of the degree, **Computer Science Engineering (Artificial Intelligence and Machine Learning)**.

Aditya Sajja(12107787)

Signature of the student



Dated:

PROJECT COMPLETION CERTIFICATE



PROJECT COMPLETION CERTIFICATE

In recognition of the commitment to achieve professional excellence this is
to certify that Ms./Mr.

Aditya Sajja

has successfully completed an Industry-oriented project.

Project Name _____ Car Price Prediction
Technologies Used _____ Data Science and Machine Learning- Anaconda, Jupyter Notebook
Reference No. _____ MO/CEP2023/E/ 2211
Training Date _____ June, 2023 – July, 2023
Training Duration _____ 6 Weeks
Training Location _____ Allsoft Solutions & Services Private Limited


Program Co-ordinator
Industry/Academic Alliance




Director
Training and Development
Allsoft Solutions and Services

BIG DATA - ANALYTICS IoT ORACLE J2EE PHP CLOUD COMPUTING

DECLARATION LETTER



A Pioneer organization & IBM Business Partner

Date: July, 2023

TO WHOM IT MAY CONCERN

This is to certify, Aditya Sajja student of Lovely Professional University, Phagwara has undergone 6 Weeks Summer Training on IBM project and technologies with us. The details are as follows: -

PROJECT NAME	Car Price Prediction
TRAINING PERIOD	June, 2023 - July, 2023
TECHNOLOGY	Data Science and Machine Learning- Anaconda, Jupyter Notebook
DURATION OF TRAINING	6 Weeks
REFERENCE NUMBER	MO/CEP2023/E/ 2211
SUBJECT MATTER EXPERT	Mr. Mayank Raghuvanshi
ACHIEVEMENTS	Project Completion Certificate and Declaration Letter

During the training, assessment and project period we find the students sincere, hardworking and having good behavior and moral character.

We wish intern all success in future endeavors.

Mr. S. K Garg
In charge | Delivery
Allsoft Solutions and Services




For Allsoft Solutions & Services
Authorised Signatory

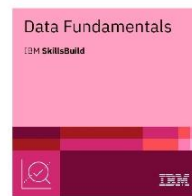
IBM SKILLS BUILD CERTIFICATE



Data Fundamentals

ISSUED TO

Aditya Sajja



Issued on: 20 JUL 2023 | Issued by: IBM-SkillsBuild
Verify: <https://www.credly.com/go/ITZeo0K3>

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude and appreciation to IBM for providing me with the opportunity to undertake a transformative summer internship on the fascinating intersection of machine learning and data science. This experience has been invaluable in enhancing my knowledge and skills in the field, and I am deeply grateful for the guidance, mentorship, and resources that were extended to me throughout my journey.

I am indebted to my supervisor and the entire team for their unwavering support and insightful feedback, which greatly contributed to my growth during this internship. The collaborative and innovative environment at IBM has been truly inspiring, allowing me to engage with cutting-edge technologies and real-world challenges.

Furthermore, I want to extend my gratitude to my fellow interns, whose camaraderie and shared enthusiasm added an enriching dimension to my internship experience.

Lastly, I would like to thank my family and friends for their continuous encouragement and belief in my abilities. This internship has opened new doors of understanding and has set me on a path to contribute meaningfully to the ever-evolving landscape of machine learning and data science.

With profound appreciation,

Aditya Sajja

LIST OF CONTENTS

S. No.	Title	Page
1	Declaration by Student	2
2	Project Completion Certificate	3
3	Declaration Letter	4
4	IBM Skills build Certificate	5
5	Acknowledgement	6
6	List of Contents	7
7	Chapter-1 Introduction to Data Science	8
8	Predictive Modelling	16
9	Weekly Overview of Course activities	32
10	Chapter-2 Work done during Summer Internship	35
11	Chapter-3 Conclusion	45
12	References	47

INTRODUCTION TO DATA SCIENCE

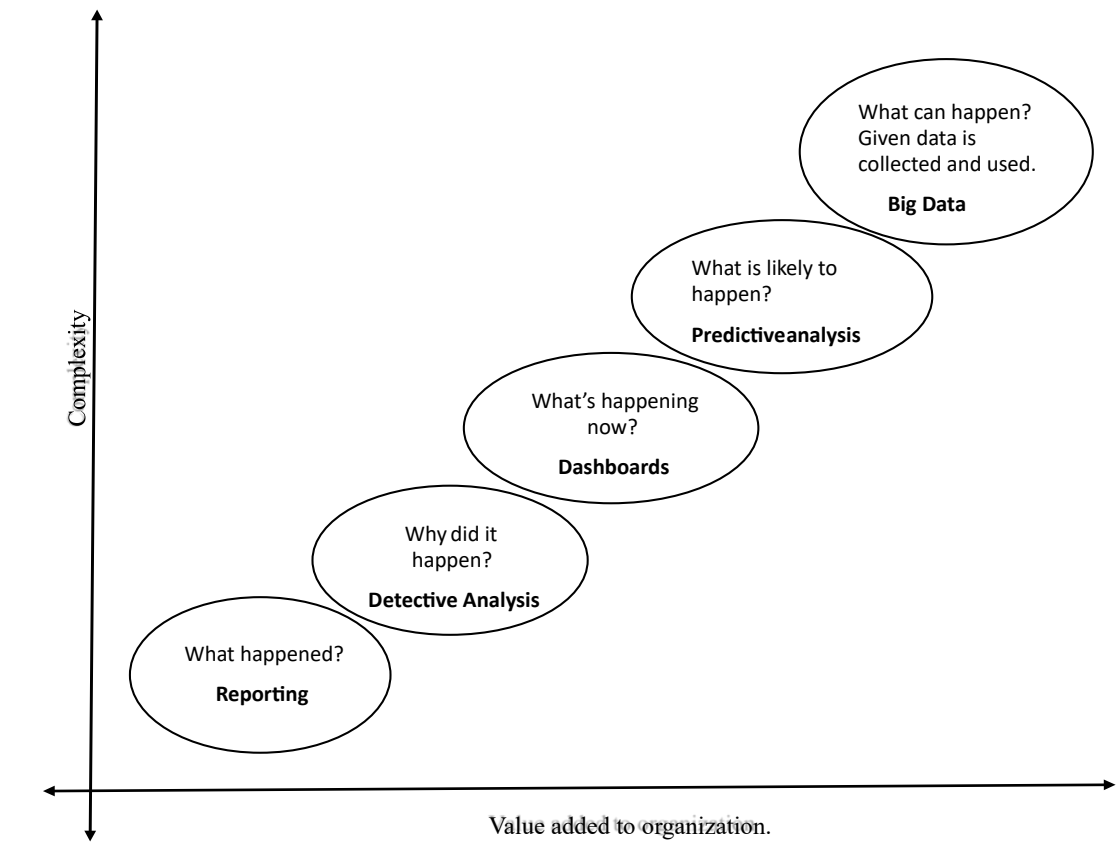
Data Science:

The field of bringing insights from data using scientific techniques is called **data science**.

Amazon Go – No checkout line.

Computer Vision - The advancement in recognizing an image by a computer involves processing large sets of image data from multiple objects of the same category. For example, Face recognition.

Spectrum of Business Analysis:



Reporting / Management Information System

To track what is happening in an organization.

Detective Analysis

Asking questions based on data we are seeing, like. Why did something happen?

Dashboard / Business Intelligence

Utopia of reporting. Every action about business is reflected in front of the screen.

Predictive Modelling

Using past data to predict what is happening at granular level.

Big Data

Stage where complexity of handling data gets beyond the traditional system.

Can be caused because of volume, variety, or velocity of data. Use specific tools to analyze such scale data.

Applications of Data Science

- **Recommendation System**

Example-In Amazon recommendations are different for different users according to their past search.

- **Social Media**

1. Recommendation Engine

2. Ad placement

3. Sentiment Analysis

- Deciding the right credit limit for credit card customers.

- Suggesting right products from e-commerce companies

1. Recommendation System

2. Past Data Searched

3. Discount Price Optimization

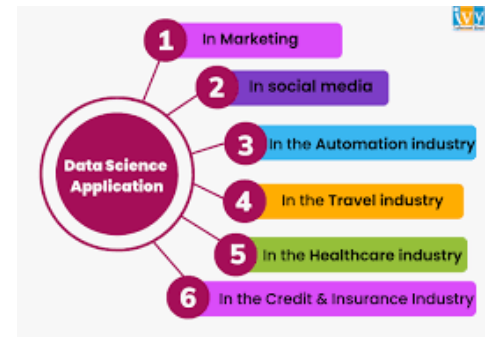
- How google and other search engines know what are the more relevant results for our search query?

1. Apply ML and Data Science

2. Fraud Detection

3. AD placement

4. Personalized search results



Python Introduction

Python is a general-purpose, interpreted programming language. Its object-oriented programming methodology is straightforward but efficient, and it includes good high-level data structures. Python is a fantastic language for scripting and quick application development in many domains on most platforms because to its clean syntax, dynamic typing, and nature of being an interpreted language.

Python for Data science:

Why Python???

1. Python is an open-source language.
 2. Syntax as simple as English.
 3. Very large and Collaborative developer community.
 4. Extensive Packages.
- UNDERSTANDING OPERATORS:
 - Theory of operators: - Operators are symbolic representation of Mathematical tasks.
 - VARIABLES AND DATATYPES:
 - Variables are named bound to objects. Data types in python are int (Integer), Float, Boolean and strings.
 - CONDITIONAL STATEMENTS:
 - If-else statements (Single condition)
 - If- elif- else statements (Multiple Condition)
 - FUNCTIONS:
 - Functions are re-usable pieces of code. Created for solving specific problems.
 - Two types: Built-in functions and User- defined functions.

- Functions cannot be reused in python.
- **LISTS:** A list is an ordered data structure with elements separated by comma and enclosed within square brackets.
- **DICTIONARY:** A dictionary is an unordered data structure with elements separated by comma and stored as key: value pair, enclosed with curly braces {}.

Statistics:

Descriptive Statistic

Mode:

It is a number which occurs most frequently in the data series.

It is robust and is not generally affected much by the addition of a couple of new values.

Code

```
import pandas as pd
data=pd.read_csv( "Mode.csv")      //reads data from csv file
data.head()                       //print first five lines
mode_data=data['Subject'].mode()   //to take mode of subject column
print(mode_data)
```

Mean:

```
import pandas as pd
data=pd.read_csv( "mean.csv")      //reads data from csv file
data.head()                       //print first five lines
mean_data=data[Overallmarks].mean() //to take mode of subject column
print(mean_data)
```

Median:

Absolute central value of data set.

```
import pandas as pd
data=pd.read_csv( "data.csv")           //reads data from csv file
data.head()                             //print first five lines
median_data=data[Overallmarks].median() //to take mode of subject
column print(median_data)
```

Types of Variables:

- Continuous – Which takes continuous numeric values. E.g.-marks
- Categorical-Which has discrete values. E.g.- Gender
- Ordinal – Ordered categorical variables. E.g.- Teacher feedback
- Nominal – Unordered categorical variable. E.g.- Gender

Outliers:

Any value which will fall outside the range of the data is termed as a outlier. E.g.- 9700 instead of 97.

Reasons of Outliers:

- Typos-During collection. E.g.-adding extra zero by mistake.
- Measurement Error-Outliers in data due to measurement operator being faulty.
- Intentional Error-Errors which are induced intentionally. E.g.- claiming smaller amount of alcohol consumed then actual.
- Legit Outlier—These are values which are not actually errors but in data due to legitimate reasons. E.g. - a CEO's salary might be high as compared to other employees.

Interquartile Range (IQR):

Is the difference between third and first quartile from last. It is robust to outliers.

Histograms:

Histograms depict the underlying frequency of a set of discrete or continuous data that are measured on an interval scale.

```
import pandas as pd
histogram=pd.read_csv(histogram.csv)
import matplotlib.pyplot as plt
%matplotlib inline
plt.hist(x= 'Overall Marks',data=histogram)
plt.show()
```

Inferential Statistics:

Inferential statistics allow us to make inferences about the population from the sample data.

Hypothesis Testing:

Hypothesis testing is a kind of statistical inference that involves asking a question, collecting data, and then examining what the data tells us about how to proceed. The hypothesis to be tested is called the null hypothesis and given the symbol H_0 . We test the null hypothesis against an alternative hypothesis, which is given the symbol H_a .

Decision Made	Null Hypothesis is True	Null Hypothesis is False
Reject Null Hypothesis	Type I Error	Correct Decision
Do not Reject Null Hypothesis	Correct Decision	Type II Error

T Tests:

When we have just a sample not population statistics.

Use sample standard deviation to estimate population standard deviation.

T test is more prone to errors because we just have samples.

Z Score:

The distance in terms of number of standard deviations, the observed value is away from mean, is standard score or z score.

$$Z = \frac{\bar{X} - \mu}{\sigma}$$

+Z – value is above mean.

-Z – value is below mean.

The distribution once converted to z- score is always same as that of shape of original distribution.

Chi Squared Test:

To test categorical variables.

Correlation:

Determine the relationship between two variables.

It is denoted by r. The value ranges from -1

to +1. Hence, 0 means no relation.

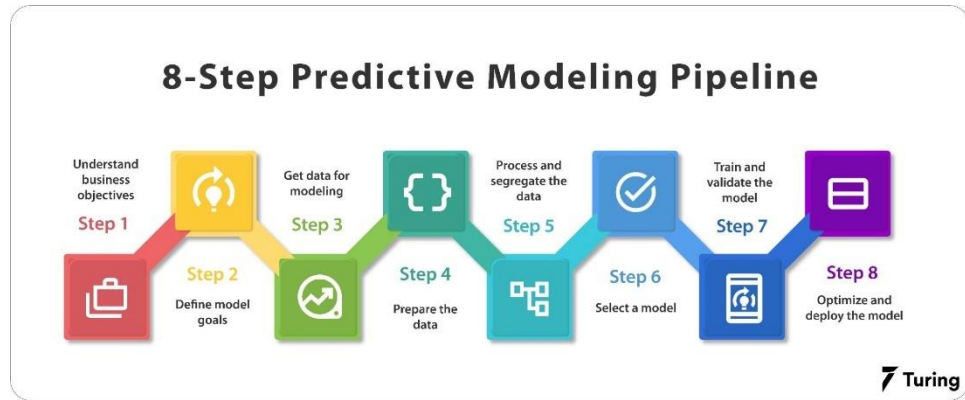
Syntax

```
import pandas as pd
import numpy as np
data=pd.read_csv("data.csv")
data.corr()
```

Predictive Modelling:

Making use of past data and attributes we predict future using this data.

Eg-



Past	Horror Movies
Future	Unwatched Horror Movies

Predicting stock price movement:

1. Analysing past stock prices.
2. Analysing similar stocks.
3. Future stock price required.

Types:

1. Supervised Learning

Supervised learning is a type of algorithm that uses a known dataset (called the training dataset) to make predictions. The training dataset includes input data and response values.

- **Regression**-which has continuous possible values. Eg-Marks
- **Classification**-which has only two values. Eg-Cancer prediction is either 0 or 1.

2. Unsupervised Learning

Unsupervised learning is the training of machine using information that is neither classified nor. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

- **Clustering:** A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association:** An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Stages of Predictive Modelling:

1. Problem definition
2. Hypothesis Generation
3. Data Extraction/Collection
4. Data Exploration and Transformation
5. Predictive Modelling
6. Model Development/Implementation

Problem Definition:

Identify the right problem statement, ideally formulate the problem mathematically.

Hypothesis Generation:

List down all possible variables, which might influence the problem objective. These variables should be free from personal bias and preferences.

Quality of model is directly proportional to quality of hypothesis.

Data Extraction/Collection:

Collect data from different sources and combine those for exploration and model building.

While looking at data we might come across a new hypothesis.

Data Exploration and Transformation:

Data extraction is a process that involves retrieval of data from various sources for further data processing or data storage.

Steps of Data Extraction

- Reading the data Eg- From csv file
- Variable identification
- Univariate Analysis
- Bivariate Analysis
- Missing value treatment
- Outlier treatment
- Variable Transformation

Variable Treatment

It is the process of identifying whether variable is:

1. Independent or dependent variable
2. Continuous or categorical variable

Why do we perform variable identification?

1. Techniques like supervised learning require identification of dependent variables.
2. Different data processing techniques for categorical and continuous data.

Categorical variable- Stored as object.

Continuous variable-Stored as int or float.

Univariate Analysis:

1. Explore one variable at a time.
2. Summarize the variable.
3. Make sense out of that summary to discover insights, anomalies, etc.

Bivariate Analysis:

- When two variables are studied together for their empirical relationship.
- When you want to see whether the two variables are associated with each other.
- It helps in prediction and detecting anomalies.

Missing Value Treatment:

1. Non-response – E.g.-when you collect data on people's income, and many choose not to answer.
2. Error in data collection. E.g.- Faculty data
3. Error in data reading.

Types:

1. MCAR (Missing completely at random): Missing values have no relation to the variable in which missing value exist and other variables in dataset.
2. MAR (Missing at random): Missing values have no relation to the in which missing value exist and the variables other than the variables in which missing values exist.

3. MNAR (Missing not at random): Missing values have relation to the variable in which missing value exists.

Identifying:

Syntax: -

1. describe() :- gives statistical analysis.
2. Isnull() :- Output will be in True or False

Different methods to deal with missing values:

1. Imputation :-

Continuous-Impute with help of mean, median or regression mode.

Categorical-With mode, classification model.

2. Deletion :-

Row wise or column wise deletion. But it leads to loss of data.

Outlier Treatment:

Reasons of Outliers:

1. Data entry Errors
2. Measurement Errors
3. Processing Errors
4. Change in underlying population.

Types of Outliers:

Univariate

Analyzing only one variable for outlier.

Eg – In box plot of height and weight.

Weight will be analyzed for outliers.

Bivariate

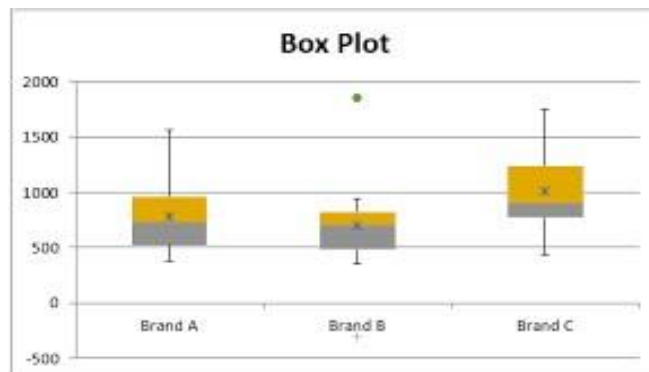
Analyzing both variables for outlier.

Eg- In scatter plot graph of height and weight. Both will be analyzed.

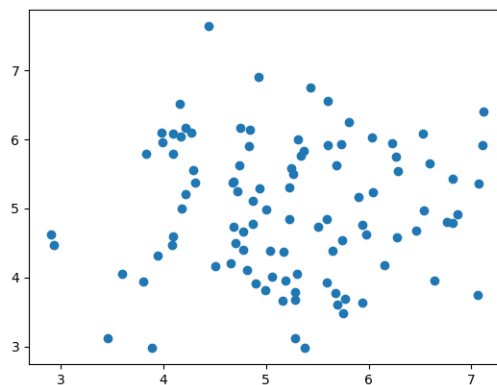
Identifying Outlier

Graphical Method

- Box Plot :



- Scatter Plot :



Formula Method

Using Box Plot

$< Q1 - 1.5 * IQR$ or $> Q3 + 1.5 * IQR$

Where $IQR = Q3 - Q1$

$Q3 = \text{Value of 3}^{\text{rd}} \text{ quartile}$

Q1=Value of 1st quartile

Treating Outliers

1. Deleting observations.
2. Transforming and binning values.
3. Imputing outliers is like missing values.
4. Treat them as separate.

Variable transformation

Is the process in which:

1. We replace a variable with some function of that variable. Eg – Replacing a variable x with its log.
2. We change the distribution or relationship of a variable with others.

Used to –

1. Change the scale of a variable
2. Transforming non-linear relationships into linear relationship
3. Creating symmetric distribution from skewed distribution.

Common methods of Variable Transformation – Logarithm, Square root, Cube root, Binning, etc.

Model Building:

It is a process to create a mathematical model for estimating / predicting the future based on past data.

E.g.

A retailer wants to know the default behavior of its credit card customers.

They want to predict the probability of default for each customer in the next three months.

- The probability of default would lie between 0 and 1.

- Assume every customer has a 10% default rate.

Probability of default for each customer in next 3 months=0.1

It moves the probability towards one of the extremes based on attributes of past information.

A customer with volatile income is more likely (closer to) to default.

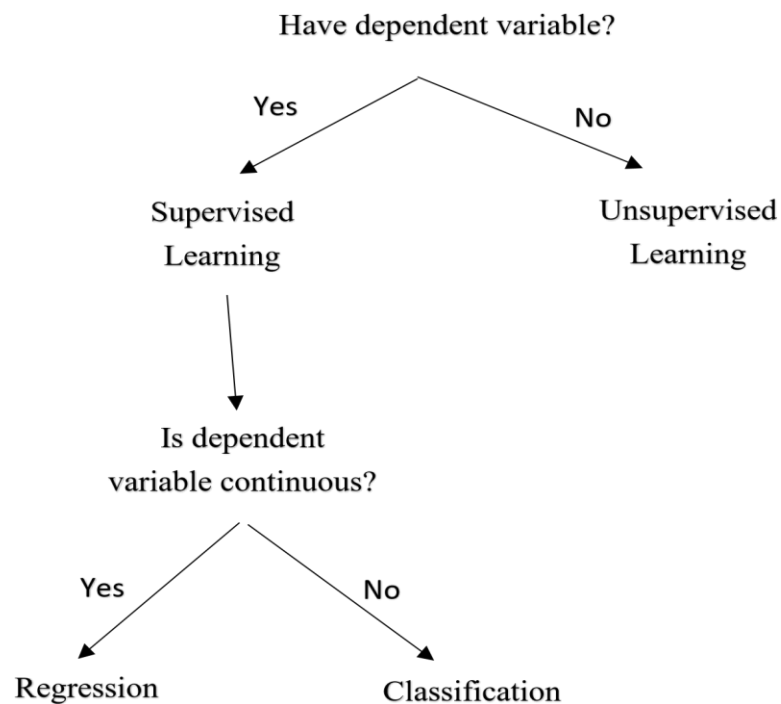
A customer with a healthy credit history for the last few years has low chances of default (closer to 0).

Steps in Model Building:

1. Algorithm Selection
2. Training Model
3. Prediction / Scoring

Algorithm Selection:

Example-



Eg- Predict whether the customer will buy a product or not.

Algorithms:

- Logistic Regression
- Decision Tree
- Random Forest

Training Model:

It is a process to learn the relationship / correlation between independent and dependent variables.

We use dependent variable of train data set to predict/estimate.

Dataset

- Train

Past data (known dependent variable).

Used to train models.

- Test

Future data (unknown dependent variable)

Used to score.

Prediction / Scoring

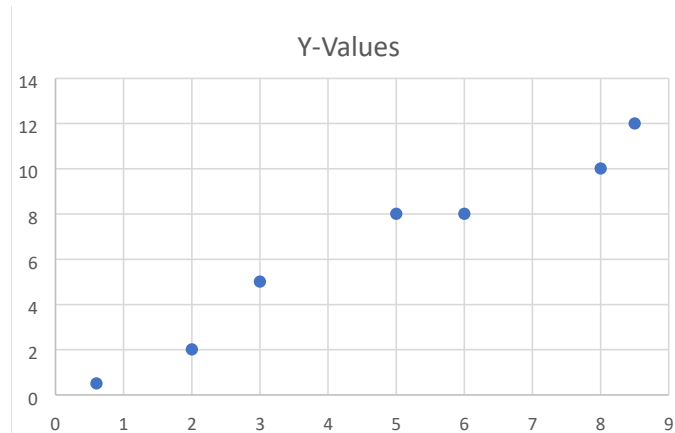
It is the process to estimate/predict dependent variable of train data set by applying model rules.

Algorithms of Machine Learning:

Linear Regression:

Linear regression is a statistical approach for modelling the relationship between a dependent variable with a given set of independent variables.

It is assumed that the two variables are linearly related. Hence, we try to find a linear function. That predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).



The equation of regression line is represented as:

$$h(x_i) = \beta_0 + \beta_1 x_i$$

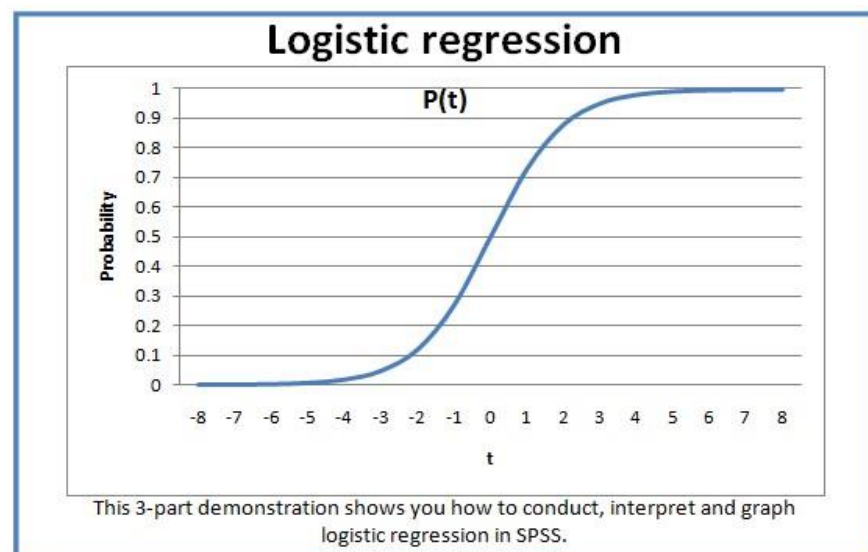
The squared error or cost function, J as:

$$J(\beta_0, \beta_1) = \frac{1}{2n} \sum_{i=1}^n$$

Logistic Regression:

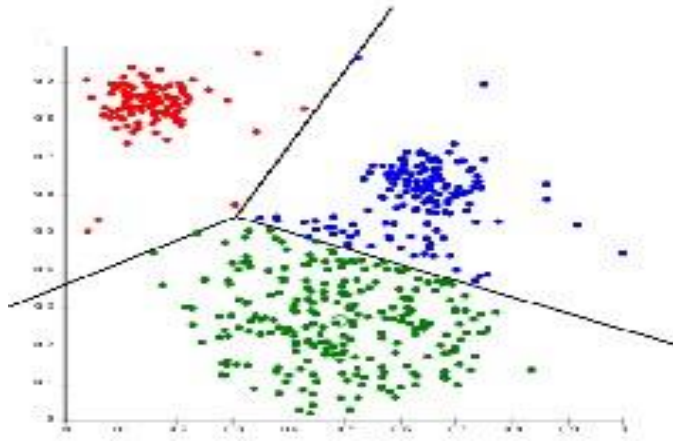
Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.

$$C = -y (\log(y) - (1-y) \log(1-y))$$



K-Means Clustering (Unsupervised learning):

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.



Introduction to Big Data:

Why is data Important?

Data is critical for business and has big value. Data is one of the most valuable assets organizations can have, whether in business, finance, healthcare, retail, technology, marketing, or other industries. The number of companies using data insights continues to grow. Data insights have the potential to help many companies:

- Improve operations.
- Better understand end users or customers.
- Drive efficiently.
- Reduce costs.
- Increase profits.
- Find new innovations.
- **Data is a problem solver.**

Data analysts spend a lot of time working in a database. A database is an organized collection of structured data in a computer system.

Transforming data into standard format (or tidy data) makes storage and analysis easier.

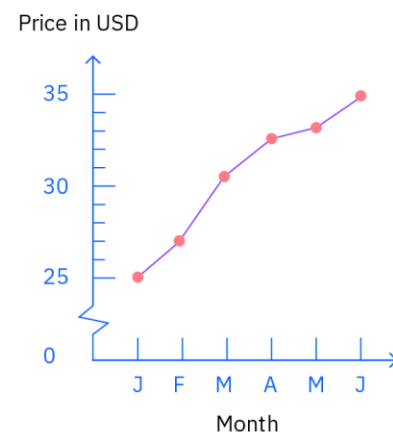
Then what is Big Data:

There is no official definition for big data, but according to tech giants Big Data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making and process automation.

Stock J



Stock J



Data can be misleading.

While each line chart presents the trending price over time for "Stock J", the vertical scale (or y-axis) for Price is different. The scales show the data in two different increments. Notice the second chart is misleading because it doesn't depict \$0 to 25 for Price like the first chart does.

And it shows Price in \$5 increments. It makes it look like "Stock J" increased in price faster! The first chart is a more accurate depiction because it does not skip the price from \$0 to \$25 and shows Price consistently in \$10 increments.

The key point here is to be precise in how you choose to depict data.

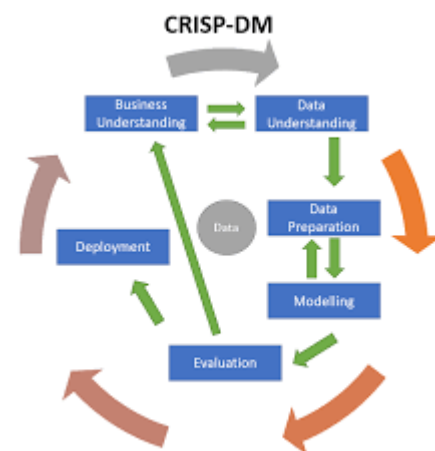
There are four types of data analytics that answer key questions, build on each other, and increase in complexity:

- Descriptive
- Diagnostic
- Predictive
- Prescriptive

There are three classic and widely adopted data science methodologies:

- CRISP-DM stands for Cross-Industry Standard Process for Data Mining. consists of six phases with arrows indicating the most important and frequent dependencies between phases:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modelling
5. Evaluation



6. Deployment

➤ KDD stands for Knowledge Discovery in Database.

1. Selection
2. Preprocessing
3. Transformation
4. Data Mining
5. Interpretation/Evaluation

➤ SEMMA stands for its five steps:

1. **S**ample
2. **E**xplore
3. **M**odify
4. **M**odel
5. **A**ssess

Overview of Data Tools and Languages:

➤ open-source industry tools. **Git** and **GitHub** are two related, but separate, platforms that are extremely popular and widely used by open-source contributors.

1. Host your own open-source project. To do this, you create an online repository and add files.
2. Contribute to an existing open-source project that's public. To do this, you access a copy of the project's repository, make updates, and request a review of the changes you want to contribute.

➤ **Structured Query Language (SQL)** is a standard language to communicate with databases. With SQL, you can:

1. Execute queries against a database.
2. Retrieve data from a database.
3. Insert records in a database.
4. Update records in a database.
5. Delete records from a database.
6. Create new databases.
7. Create new tables in a database.
8. Create stored procedures in a database.
9. Create views in a database.
10. Set permissions on tables, procedures, and views.



➤ **Python:**

1. You can use Python to connect to database systems and
2. read and modify files.
3. Python can handle big data and perform complex mathematics.
4. You can pair Python with a data manipulation and analysis software library, like pandas. Python can help you obtain insights and create data visualizations.
5. Python is very popular for data analysis also it is an open-source programming language.



- **IBM Watson Studio:** It's a collaborative data science and machine learning environment.

1. IBM Watson Studio works with open-source tools.
2. IBM Watson Studio offers a graphical interface with built-in operations.
3. You don't need to know how to code to use the tool.
4. And, IBM Watson Studio has a built-in data refinery tool.



- **Matplotlib:**

1. A Python Matplotlib script is structured so that, in most instances, a few lines of code can generate a visual data plot.
2. You can create different types of plots, such as scatterplots, histograms, bar charts, and more.
3. The visualizations can be static, animated, and interactive.
4. You can export to many different types of file formats.

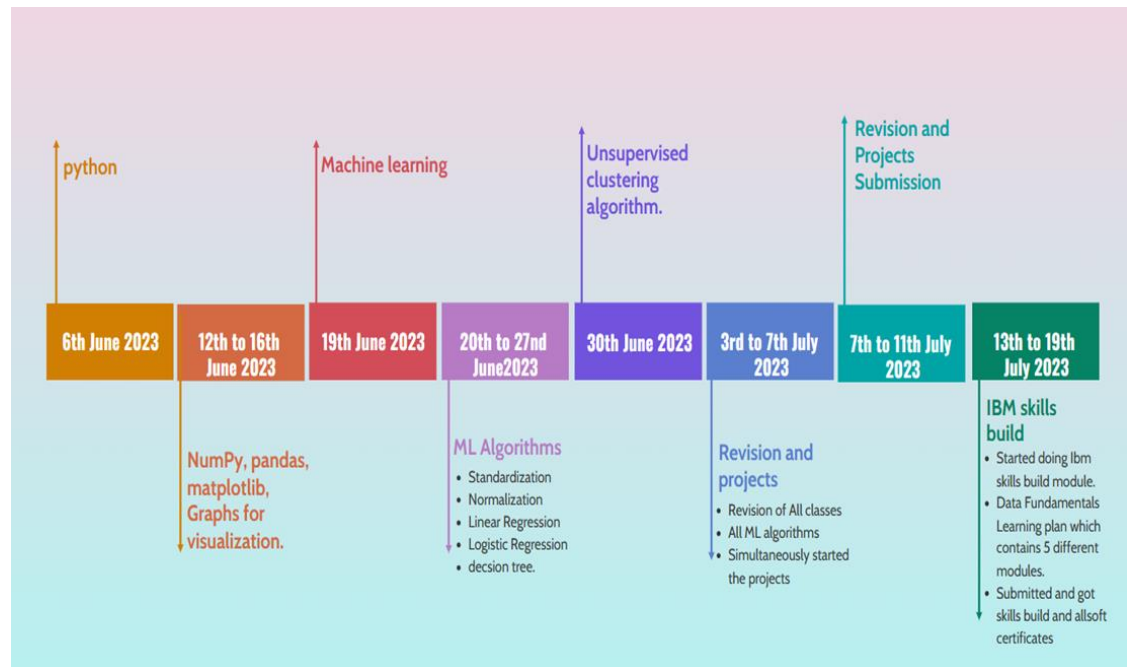


- **Google Sheets** is a free tool you can use to perform tasks like entering, analyzing, and visualizing data to make data-driven decisions.



Google Sheets

WEEKLY OVERVIEW OF COURSE ACTIVITIES



WEEK 1:

DAY 1: Introduction- Math's for Data Analytics, Basic Probability for Data Science.

DAY 2,3,4: Statistics – Introduction to statistics, Data and their types, Measures of Frequency and central tendency, Measures of Dispersion and Measures of shape.

DAY 5,6: Probability Distribution- Introduction to Probability Distribution, Discrete Uniform Distribution, Binomial Distribution, Poisson Distribution, Continuous Uniform Distribution, Exponential Distribution, Normal Distribution, Students T Distribution.

DAY 7: Inferential Statistics – Sampling Techniques, CLT, Confidence Interval, Hypothesis Testing, Z-Test, T-Test, Chi-squared Test and 1 Way ANOVA.

WEEK 2:

DAY 1,2: Python for Data Science – Introduction to Python Programming Language, Operators, Loops, Functions, Strings and OOPs.

DAY 3,4: Python Data Toolkit – Jupyter Notebook setup, Getting Started with Files, Inventory Management System with Files, Getting Started with OS.

DAY 5,6,7: Libraries for Data Analysis, Libraries for Data Visualization, Libraries for Deployment.

WEEK 3:

DAY 1,2: Data Analysis with Python – Getting started with pandas, Data Preprocessing, Data Analysis.

DAY 3,4: Car Price Prediction Data Analysis

DAY 5,6: Data Visualization on Car Dekho Dataset

DAY 7: GDP Analysis

WEEK 4:

DAY 1,2,3: Excel for Data Analysis – Introduction to Excel, Data Entry in Excel, Data Formatting & Validation, Functions in Excel, Hyperlinks & Illustrations in Excel, Pivot Table & Charts in Excel.

DAY 4: Shortcuts in Excel, Visual Basic Analysis.

DAY 5,6: Introduction to Tableau, Understanding the Parameters, Basic Plots in Tableau.

DAY 7: Designing the plots, Fundamentals of Tableau.

WEEK 5:

DAY 1,2: Introduction to Web Scraping, Wikipedia Scraper.

DAY 3: Extracting Data from links and websites.

DAY 4,5: YouTube Scraper, Stock Image Scraper, Stock Image Infinite Scroll.

DAY 6,7: SQL – What is DBMS, Introduction to MYSQL, Types of commands, creating Tables and Databases and Inserting Data

WEEK 6:

DAY 1,2: Introduction to Machine Learning & AI, How Data Science Comes into play, Linear Regression.

DAY 3,4: Multiple Linear Regression, Polynomial Regression.

DAY 5,6: Placement Assistance Program From IBM

2.WORK DONE DURING SUMMER INTERNSHIP

2.1 Data Collection and Understanding

During the initial phase of my internship, I sourced a comprehensive car price prediction dataset from Kaggle. This dataset encapsulated a wide array of attributes, including car model, year, kilometers driven, fuel type, seller type, and more. My primary goal was to understand the context of the dataset and its attributes, enabling me to navigate subsequent stages effectively. I faced challenges in comprehending the diverse attributes and their potential impact on car prices, but gradually gained clarity through close examination and mentor guidance. This data set contains information about used cars listed on the website. This data can be used for a lot of purposes such as price prediction to exemplify the use of linear regression in Machine Learning. The columns in the given dataset are as follows:

Column Name	Description
Car_Name	Name of Car sold
Year	Year in which car was bought
Selling_Price	Price at which car sold
Present_Price	Price of same car model in current year
Kms_Driven	Number of Kilometers Car driven before it is sold
Fuel_Type	Type of fuel Car uses
Seller_Type	Type of seller
Transmission	Gear transmission of the car (Automatic/Manual)
Owner	Number of previous owners

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

	Year	Selling_Price	Present_Price	Kms_Driven	Owner
count	301.000000	301.000000	301.000000	301.000000	301.000000
mean	2013.627907	4.661296	7.628472	36947.205980	0.039867
std	2.891554	5.082812	8.644115	38886.883882	0.212302
min	2003.000000	0.100000	0.320000	500.000000	0.000000
25%	2012.000000	0.900000	1.200000	15000.000000	0.000000
50%	2014.000000	3.600000	6.400000	32000.000000	0.000000
75%	2016.000000	6.000000	9.900000	48767.000000	0.000000
max	2018.000000	35.000000	92.600000	500000.000000	2.000000

	Car_Name	Fuel_Type	Seller_Type	Transmission
count	301	301	301	301
unique	98	3	2	2
top	city	Petrol	Dealer	Manual
freq	26	239	195	261

2.2 Data Preprocessing and Cleaning

The dataset exhibited no imperfections common to real-world data, such as missing values, duplicates, and inconsistencies. This stage required strategic decision-making to strike a balance between retaining valuable data and eliminating anomalies.

2.3 Feature Engineering

The dataset's original attributes provided a foundation, but I recognized the potential for enhancing predictive power through feature engineering. I created new features derived from existing ones, such as calculating the age of the car (To Calculate how old the car is, I created new feature "No_of_Years"). By introducing these engineered features, I aimed to infuse additional information into the models, potentially improving their accuracy.

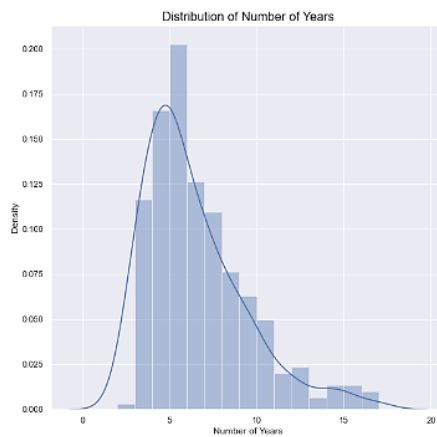
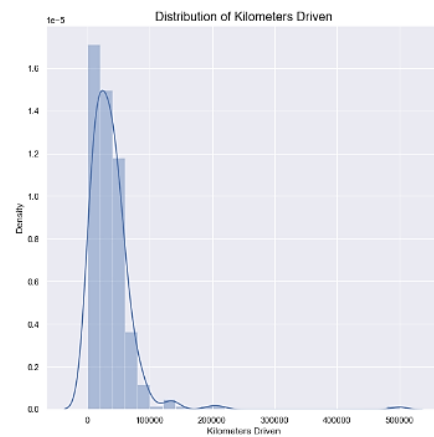
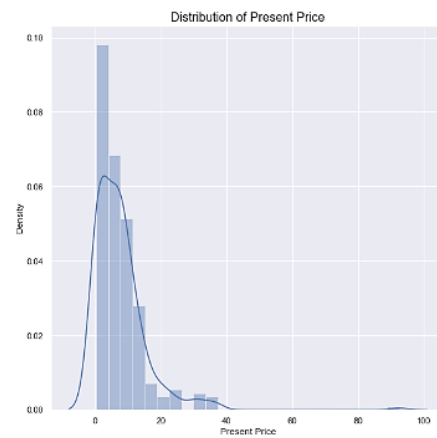
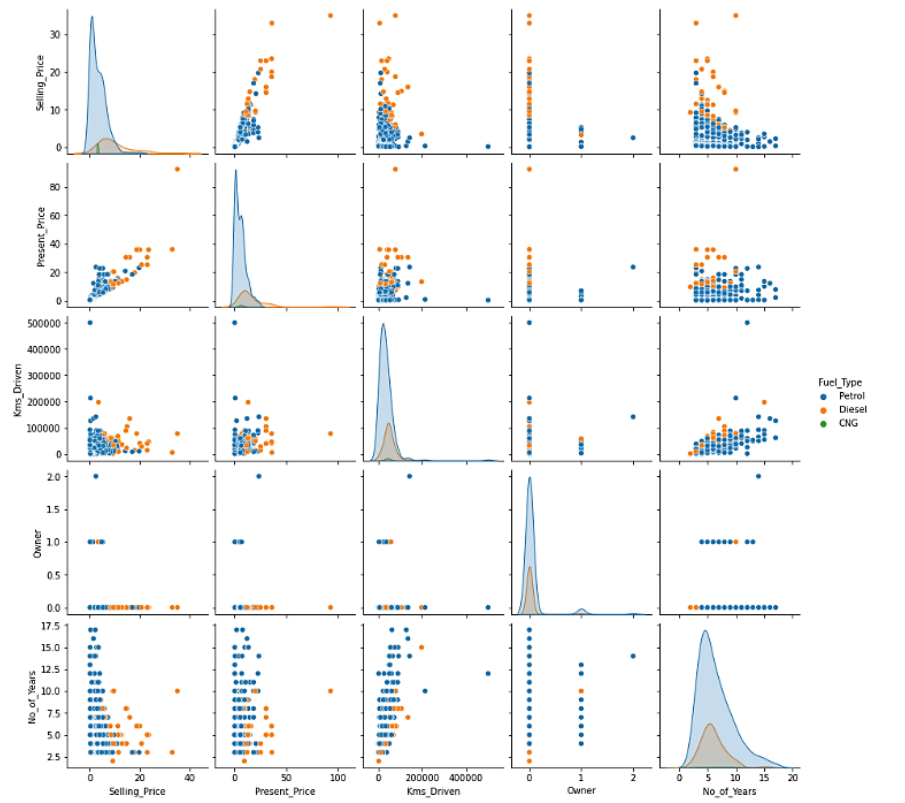
	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner	Current_Year	No_of_Years
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0	2020	6
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0	2020	7
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0	2020	3
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0	2020	9
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0	2020	6

And then I dropped the unnecessary features.

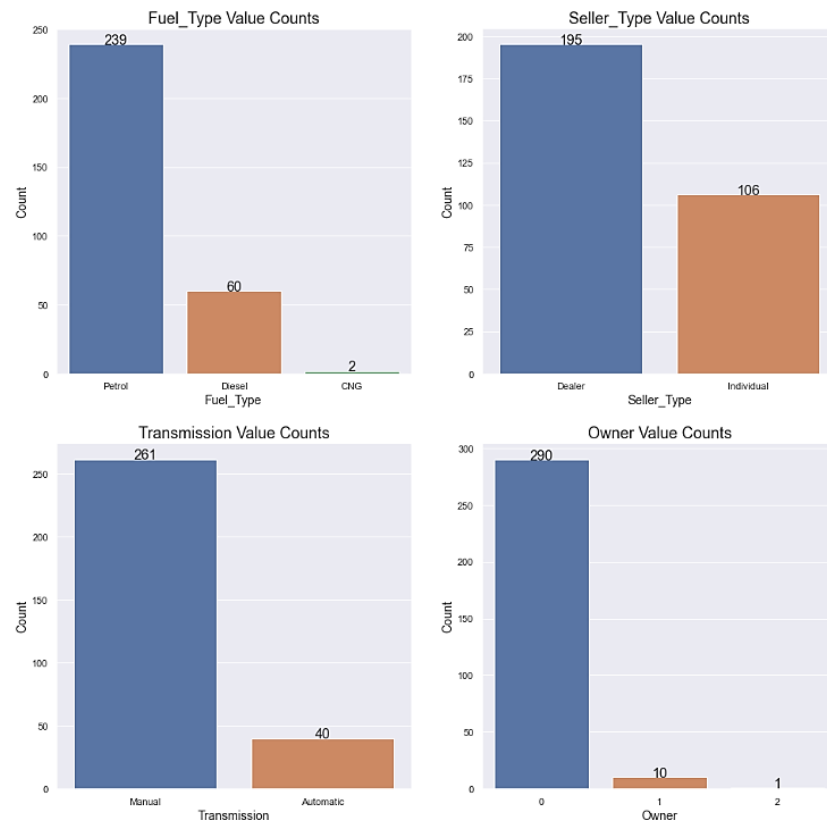
	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner	No_of_Years
0	3.35	5.59	27000	Petrol	Dealer	Manual	0	6
1	4.75	9.54	43000	Diesel	Dealer	Manual	0	7
2	7.25	9.85	6900	Petrol	Dealer	Manual	0	3
3	2.85	4.15	5200	Petrol	Dealer	Manual	0	9
4	4.60	6.87	42450	Diesel	Dealer	Manual	0	6

2.4 Exploratory Data Analysis (EDA)

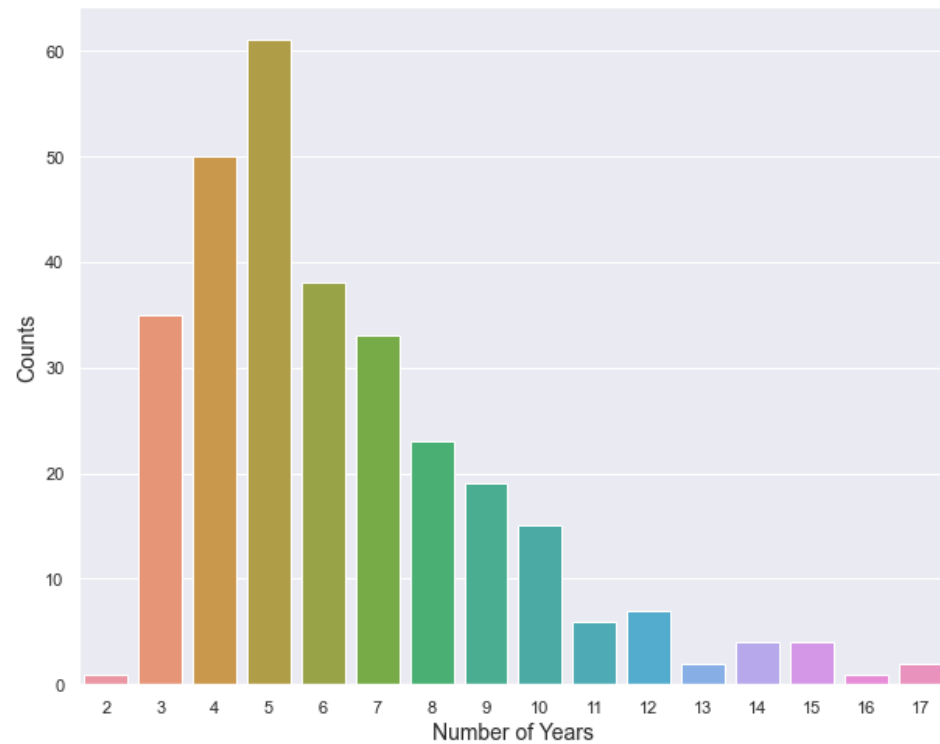
EDA was a cornerstone of my project, enabling me to gain insights into the dataset's underlying dynamics. I employed a range of visualizations, including scatter plots, histograms, and correlation matrices, to unravel relationships between attributes and the target variable—car prices. The visualizations revealed intriguing patterns, such as positive correlations between certain attributes and prices, as well as intriguing trends related to specific car makes and models.

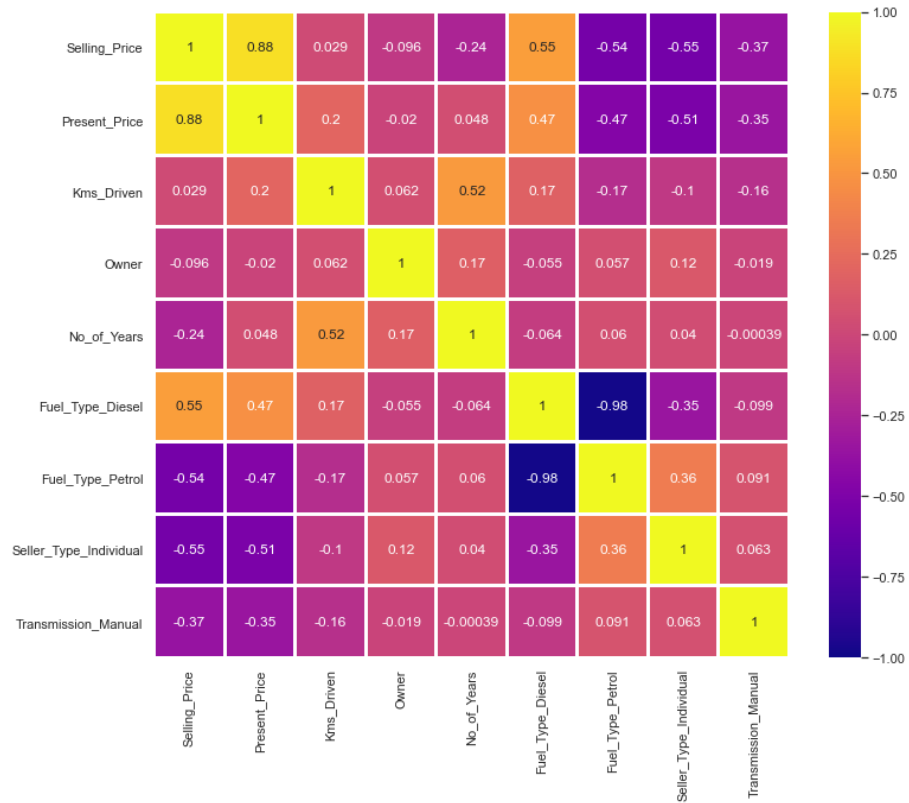


Categorical features value counts



Number of Years Value Counts





2.5 Checking all the variables

During the initial stages of my internship, a pivotal step was undertaken to ensure the robustness of the subsequent analysis. Recognizing the significance of data integrity, I meticulously scrutinized the datatypes of each variable within the dataset. This preemptive measure was driven by the intention to avert potential obstacles and errors that could potentially arise during the subsequent phases of the project. By aligning the inherent datatypes of each variable with its conceptual nature, I aimed to preemptively mitigate any mismatches or inconsistencies that might undermine the smooth progression of analyses and modeling endeavors. This proactive approach served as a foundation for a seamless workflow, fortifying the structural integrity of the dataset and laying the groundwork for accurate and unimpeded exploration and interpretation in the forthcoming stages of my internship project.


```
Selling_Price          float64
Present_Price          float64
Kms_Driven             int64
Owner                  int64
No_of_Years            int64
Fuel_Type_Diesel       uint8
Fuel_Type_Petrol       uint8
Seller_Type_Individual uint8
Transmission_Manual    uint8
dtype: object
```

2.6 Model Selection and Hyperparameter Tuning

With the dataset prepped through datatype checks and enriched via feature engineering, I embarked on the pivotal task of model selection. This phase necessitated a comprehensive evaluation of diverse machine learning algorithms tailored for regression tasks. The models under consideration ranged from the straightforward linear regression, renowned for its simplicity, to decision trees, esteemed for their interpretability, and advanced ensemble methods like random forests and gradient boosting, lauded for their heightened predictive capabilities.

Following the selection of models, the subsequent phase of hyperparameter tuning transpired. This meticulous endeavor encompassed the calibration of each model's hyperparameters—a process that entailed iterative adjustments based on cross-validation outcomes. These calibrated hyperparameters were instrumental in optimizing the models' performance, thus facilitating the extraction of meaningful insights from the dataset.

2.7 Model Training and Evaluation

With meticulously curated datasets and finely tuned models at my disposal, the dataset was partitioned into training and testing sets. Model

training commenced—a process wherein the algorithms ingested the training data to discern intricate patterns and relationships. Subsequently, the models were evaluated using a battery of robust metrics, including the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Experienced Variance Score (EVS) and R-squared.

This evaluative phase cast a spotlight on the predictive capabilities of the models, offering a glimpse into their practical applicability in real-world scenarios. The calculated metrics furnished quantitative insights into the models' performance, furnishing a foundation for informed decisions pertaining to their potential deployment.

2.8 Model Comparison and Interpretation

The culmination of the evaluation metrics from each model culminated in the process of systematic model comparison. This comparative analysis unveiled discernible trends, highlighting individual model strengths and opportunities for refinement. Notably, ensemble methods such as random forests and gradient boosting consistently outperformed their simpler counterparts. The interpretability intrinsic to decision trees added a valuable layer to the understanding of feature importance, thereby enriching the analysis of the dataset's intricate dynamics.

The outcomes of this comparison served as a springboard for the nuanced interpretation of the factors that substantively influence car prices. By dissecting the relative performance and efficacy of different models, I was primed to synthesize insights that extended beyond the confines of mere metrics, cultivating a profound comprehension of the driving forces underpinning the automotive market.

2.9 Deployment and Future Considerations

The conclusions drawn from my precise car price prediction models resonate across the echelons of various sectors, poised to empower car buyers, sellers, and manufacturers alike. This beckons contemplation regarding the practical deployment of the predictive models in real-world settings. Visioning the seamless integration of these models into decision-making fabric, I acknowledged the potential for advancing model efficacy through the infusion of external data sources and intricate feature engineering techniques.



Next, I explore the residuals to make sure everything was okay with the data (i.e., it is Normally distributed).



2.10 Collaboration and Communication

Throughout the internship, collaboration was essential for success.

Engaging with colleagues, mentors, and peers enabled me to leverage their expertise, seek guidance, and engage in fruitful discussions. Effective communication facilitated the exchange of ideas and ensured that my work aligned with broader organizational objectives.

3.CONCLUSION

The culmination of my summer internship journey on the intricate landscape of car price prediction has yielded profound insights, enriched understanding, and a promising trajectory for future applications. This chapter encapsulates the summation of findings, key observations, and a comprehensive perspective on the project's accomplishments, as well as outlining the avenues for future growth and application.

3.1 Summary of Findings and Key Observations

The journey through data science and machine learning methodologies unfolded with a dynamic interplay of exploration, analysis, and prediction. Through meticulous data preprocessing, exploratory data analysis, and feature engineering, I unearthed the intricate relationships that govern car prices. The calibrated models, including decision trees, random forests, and gradient boosting, emerged as effective tools for accurate predictions. The interpretability offered by decision trees provided profound insights into the factors that play a pivotal role in influencing car prices, from make and model to mileage and engine specifications.

The significance of feature engineering was underscored as it breathed new life into the predictive models. The innovation of features such as mileage-to-age ratios unveiled novel perspectives on the interplay of variables. The models' performance, evaluated through metrics like MAE, RMSE, and R-squared, illuminated their predictive efficacy and validated the robustness of the applied methodologies. The project's success lay in weaving these elements together, enabling comprehensive insights into the intricate web of factors that shape car prices.

3.2 Future Scope and Applicability

The outcomes of this summer internship project transcend academic boundaries, resonating with real-world implications. The predictive models wield the potential to revolutionize the decision-making processes of car buyers, sellers, and manufacturers alike. Dealerships can optimize pricing strategies, aligning them with market trends, while consumers can make informed choices grounded in data-driven insights. Moreover, insurance and financing industries could leverage these models to enhance risk assessment and pricing structures.

Furthermore, the journey undertaken during this internship is only a steppingstone toward a broader realm of exploration. The incorporation of more extensive datasets, integration with real-time market data, and the infusion of advanced deep learning techniques could potentially amplify the models' predictive capabilities. Collaborations with industry stakeholders can shape the models to cater to specific market segments, enriching their practical applicability.

In a rapidly evolving automotive landscape marked by technological disruptions and shifting consumer preferences, the significance of accurate car price prediction resonates profoundly. The outcomes of this internship have unveiled a realm of possibilities, promising to impact market transparency, consumer satisfaction, and industry strategies.

3.3 Acknowledgment and Reflection

The completion of this summer internship project has been a transformative journey that would not have been possible without the support and guidance of mentors, colleagues, and the organization. The learning experiences, challenges overcome, and insights gained have collectively nurtured my growth as a data scientist, equipping me with invaluable skills that extend far beyond the confines of this project.

3.4 Final Word

As I bring the curtain down on this chapter and the report, I reflect on the

multifaceted dimensions explored during the summer internship period. From data preprocessing to model selection, every step has illuminated a facet of the intricate tapestry that comprises data science and machine learning. The journey's significance echoes in the potential impact on industries, decision-makers, and markets, underscoring the transformative power of predictive analytics.

In this ever-evolving landscape, as the wheels of innovation continue to turn, the outcomes of this summer internship project stand as a testament to the potential harnessed through the synergy of data science, machine learning, and real-world applications.

REFERENCES:

- <https://www.kaggle.com/>
- <https://www.google.com/>
- <https://skillsbuild.org/>