ZETECH UNIVERSITY

DATA SCIENCE PROGRAMMIG WITH PYTHON

Lesson 2: Data science methodology

Objectives of Data science methodology

- It helps identify what type of patterns will be needed to address the question most effectively.
- Forming a concrete business or research problem.

Definition of data science Methodology

- **Data Science Methodology**

- is a systematic series of techniques that guides data scientists through a specified sequence of steps to the ideal approach to solving data science problems.

Def 2

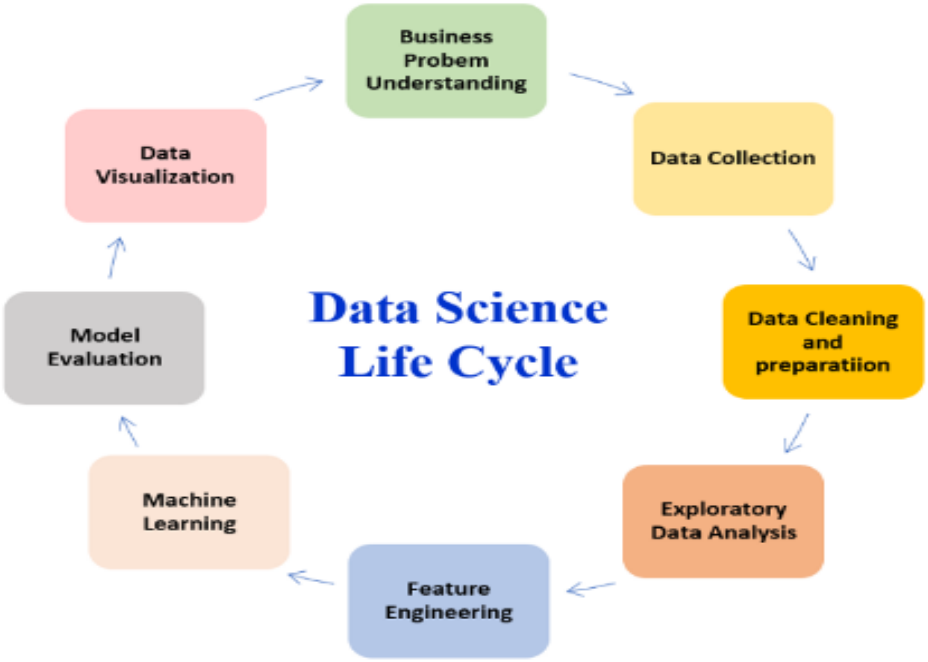
Data Science Lifecycle is a step-by-step demonstration of how machine learning and other analytical methods are used to generate insights and predictions from data to achieve a business goal.

When you enter any **computer science & IT** department on university, *generally you will learn software life cycle your first year.* And you should learn that if you want to become software developer.

So if you want to work with data or become data scientist, *you should learn about data science life cycle.*

Data Science Life Cycle Stages

1. Business Understanding and Problem Definition
2. Data Collection
3. Data Preparation /Data munging/data wrangling
4. Exploratory Data Analysis (EDA)
5. Data Modeling
6. Model Evaluation



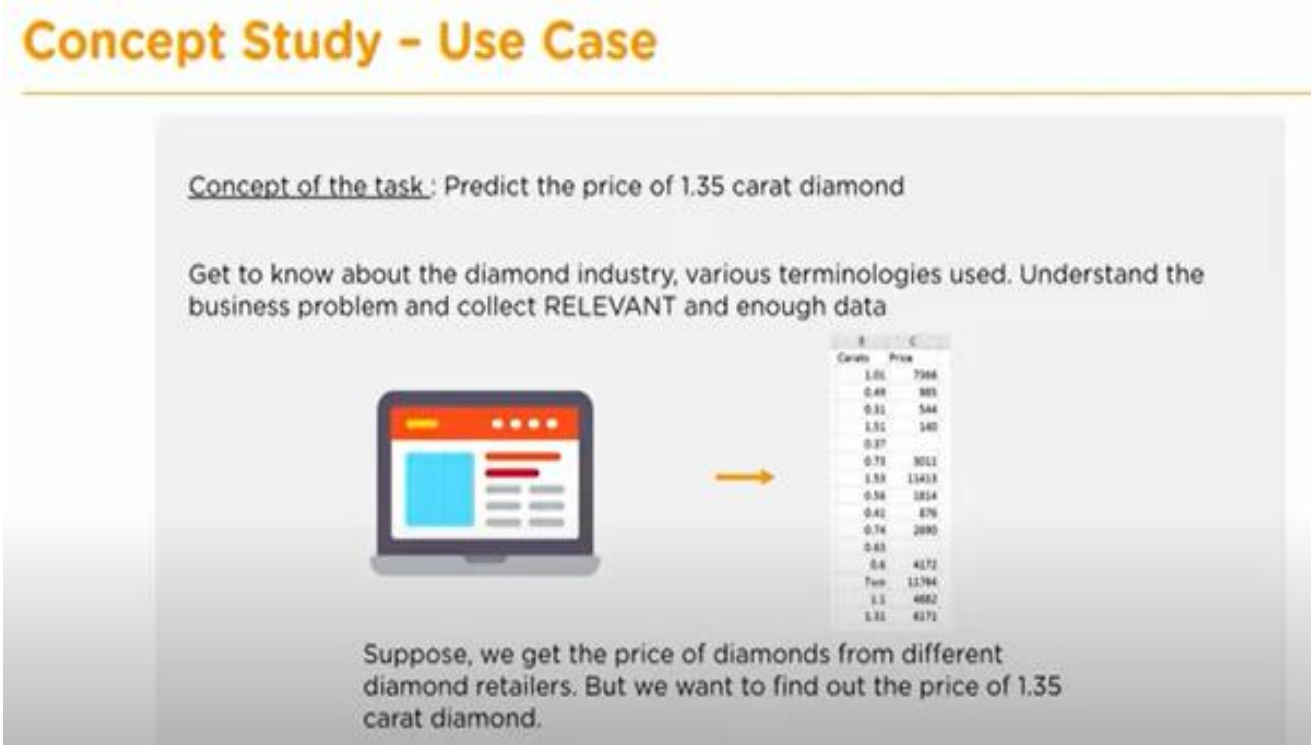
1- **Business Understanding and Problem Definition**

- Many developments in the world first started with the question of **"why".**
- It is essential to understand the business objective clearly because that will be your final goal of the analysis.

In this first phase of *data analytics*, *the stakeholders* regularly perform the following tasks examine the business trends, make case studies of similar data analytics, and study the domain of the business industry.
the stakeholders start formulating the initial hypothesis for resolving all business challenges in terms of the current market scenario.

Generally, the project lead Domain experts or product manages make that phase.

- State clearly the problem to be solved and why
- Define the potential value of the project
- Identify the project risks including ethical considerations
- Develop and communicate a high-level, flexible project plan



2. Data Collection

After gaining clarity on the problem statement, we need to collect relevant data to break the problem into small components.

The data science project starts with the identification of various

- *data sources, which may include web server logs, social media posts,* data from *digital libraries* such as the *Census datasets,* data accessed through *sources on the internet via APIs, web scraping,* or information that is *already present in an excel spreadsheet.*

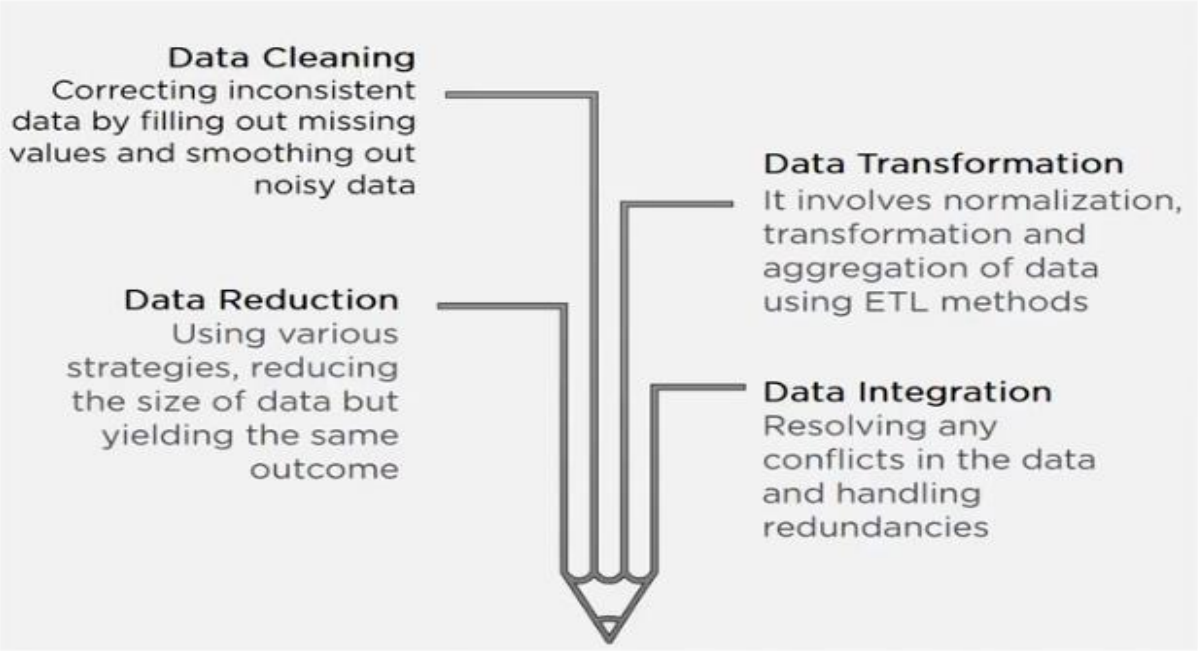
Websites that share free data sets
Kaggle.com
UCM university in US

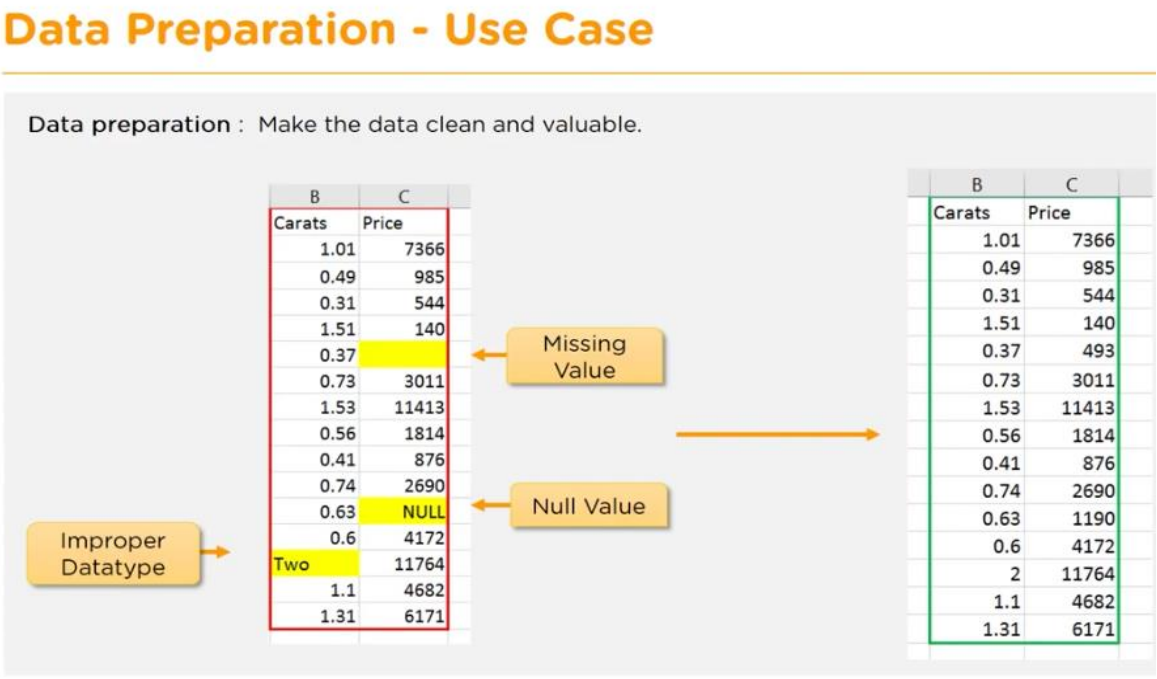
3- Data Preparation /Data munging/data wrangling

- This stage helps us gain a better understanding of the data and prepares it for further evaluation.
- **Data preparation** is the most time-consuming process, accounting for up to 90% of the total project duration**,** and this is ***the most crucial step throughout the entire life cycle.***
  - **This includes:**

- Selecting the relevant data,
- Integrating the data by merging the data sets,
- Cleaning them,
- Treating the missing values by either removing them or imputing them,
- Treating erroneous data by removing them,
- Checking outliers using box plots and handle them.

Data Preparation - Life cycle





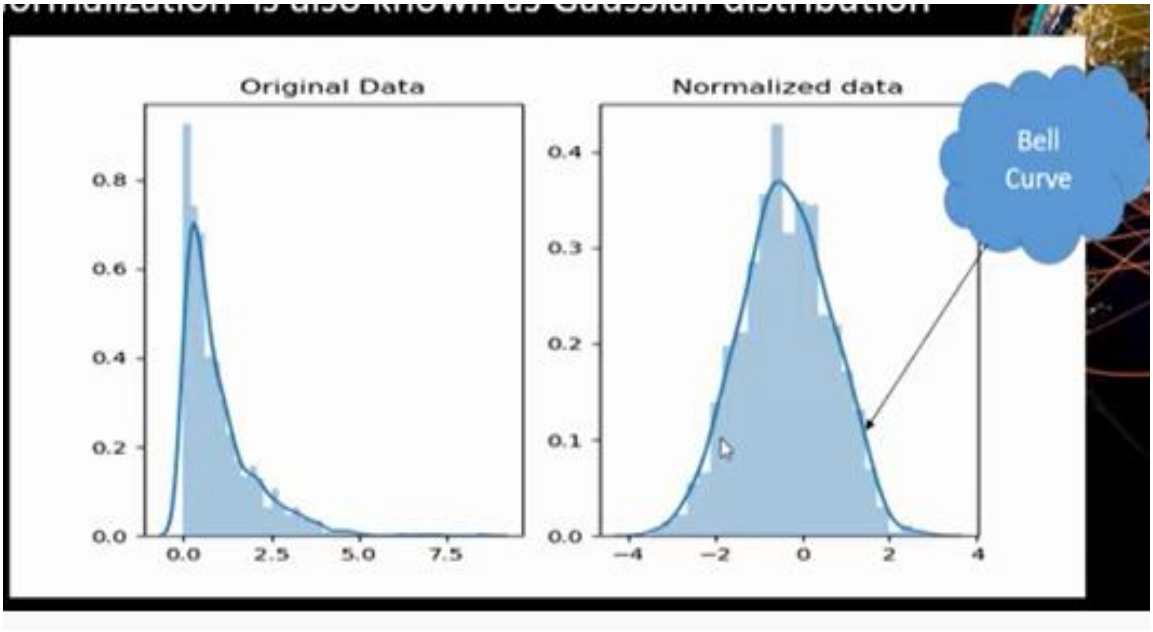
Data Preparation: - Data transformation (data format conversion)

Normalization and Aggregation using ETL methods
(ETL stands for Extract-Transform-Load. - *It is the process of moving data from multiple sources to bring it to a centralized single database*.)

What is ETL Process? (Extract Transform Load) video link
https://www.youtube.com/watch?v=g0vB5RPw8Z0&t=14s

ETL Tutorial | Extract Transform and Load video link
https://www.youtube.com/watch?v=WZw0OTgCBOY

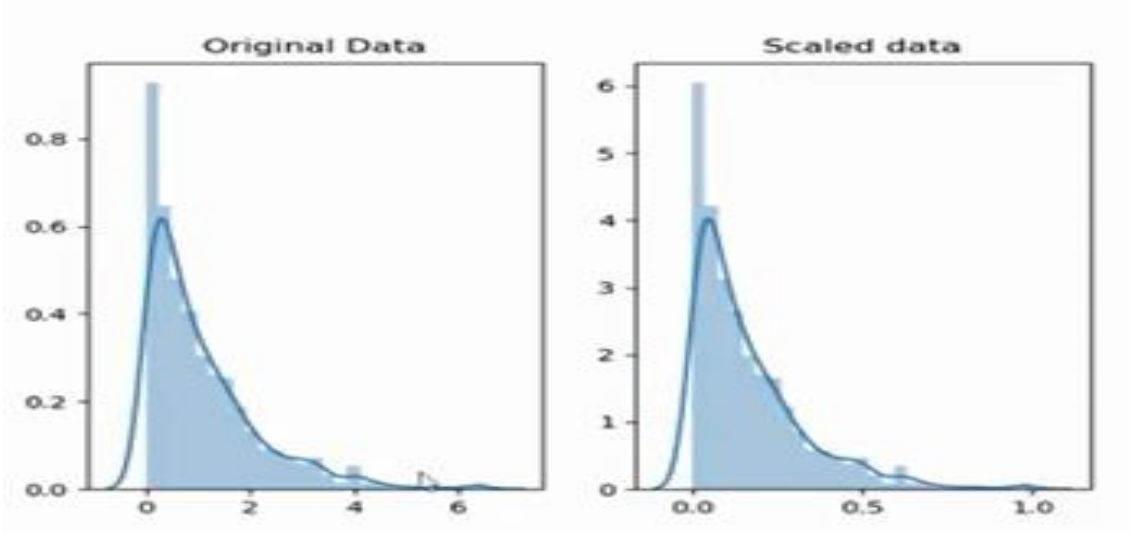
a) **Normalization** - *changing the shape of the distribution of data.*
Normalization is also known as Gaussian distribution



Normalize data for standardization purposes.

b) Scaling- *changing the range of data* **[0-1] / [0-100]**
Python libraries will do the scaling automatically

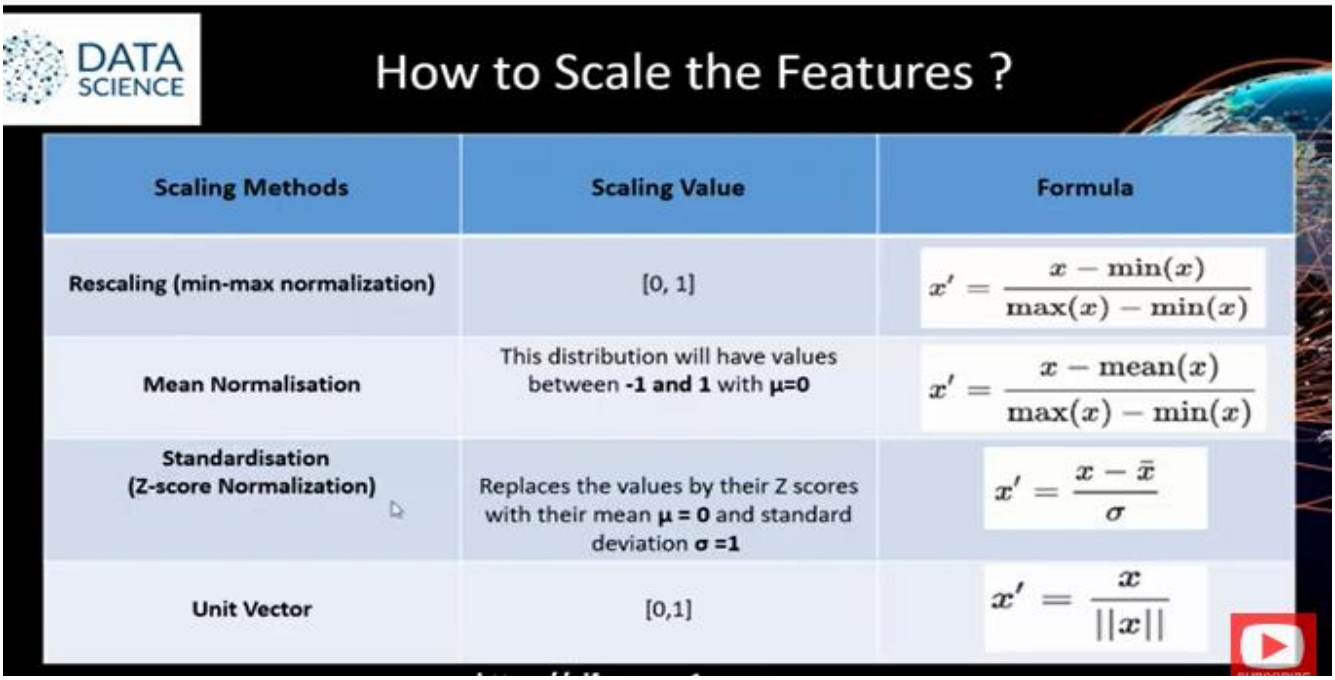
- converting all the values to be single unit.
  - Example grams' vs kilogram changed to a common unit.



Data Science Process 6: Data Cleaning Technique - Scaling and Normalization video link
https://youtu.be/O8ecZ4goHSw?t=121

Difference between Normalization and Standardization

Normalization	Standardization
This technique uses minimum and max values for scaling of model.	This technique uses mean and standard deviation for scaling of model.
It is helpful when features are of different scales.	It is helpful when the mean of a variable is set to 0 and the standard deviation is set to 1.
Scales values ranges between [0, 1] or [-1, 1].	Scale values are not restricted to a specific range.
It got affected by outliers.	It is comparatively less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for Normalization.
It is also called Scaling normalization.	It is known as Z-score normalization.
It is useful when feature distribution is unknown.	It is useful when feature distribution is normal.



https://www.youtube.com/watch?v=O8ecZ4goHSw

Data Preparation - Example

- Split the data into train data and test data in the ratio of 80:20
- It is generally advised to divide the dataset into two random partition

B	C
Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	493
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	1190
0.6	4172
2	11764
1.1	4682
1.31	6171

Train data (80%)

Test data (20%)

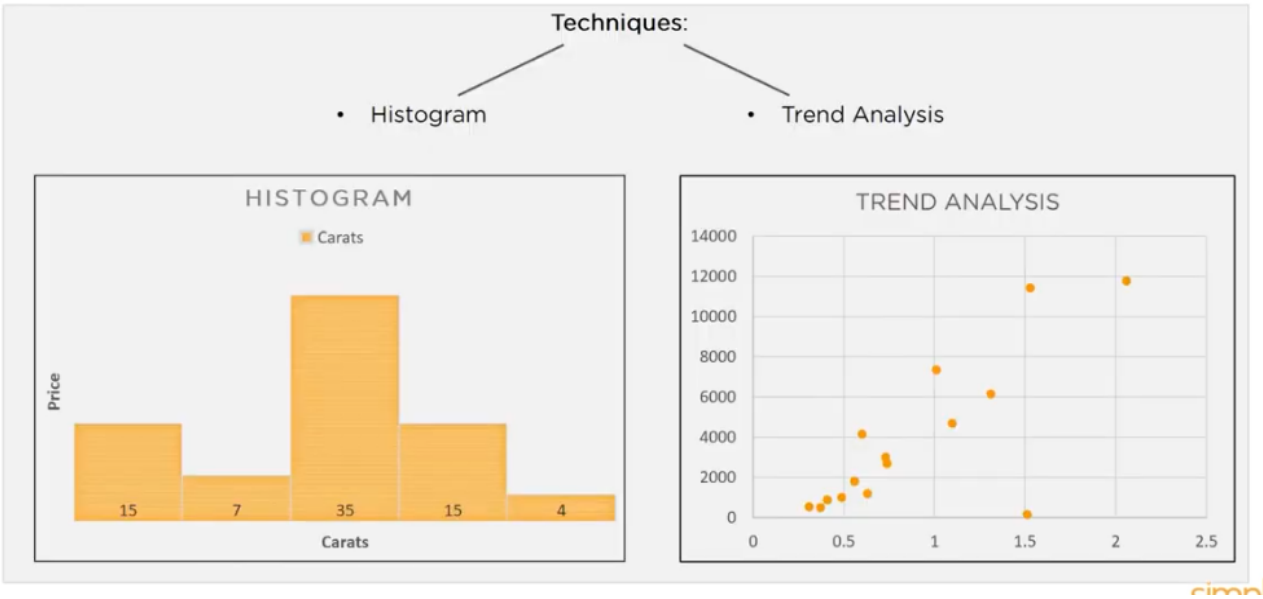
- test data set system has seen it will tend to give accurate data. but it worthwhile to test on untrained data to give correct results

Train Data - *is used to develop model*
Test Data - *is used to validate model*

4. Exploratory Data Analysis (EDA)

- An EDA is a thorough examination meant to uncover the underlying structure of a data set often with visual means and is important for a company because it
  - *exposes trends, patterns, and relationships* that are not readily apparent.
  - *Identify outliers*
  - *Understand how data is distributed*

Model Planning - Life cycle



5. Data Modeling

- most cases of data analysis, **data modeling** is regarded as the *core process*
- select the appropriate type of model that would be implemented to acquire results, whether the problem is a
  - regression problem or classification, or a clustering-based problem.

Linear regression describes the relation between 2 variables i.e. X and Y

X is Independent variable

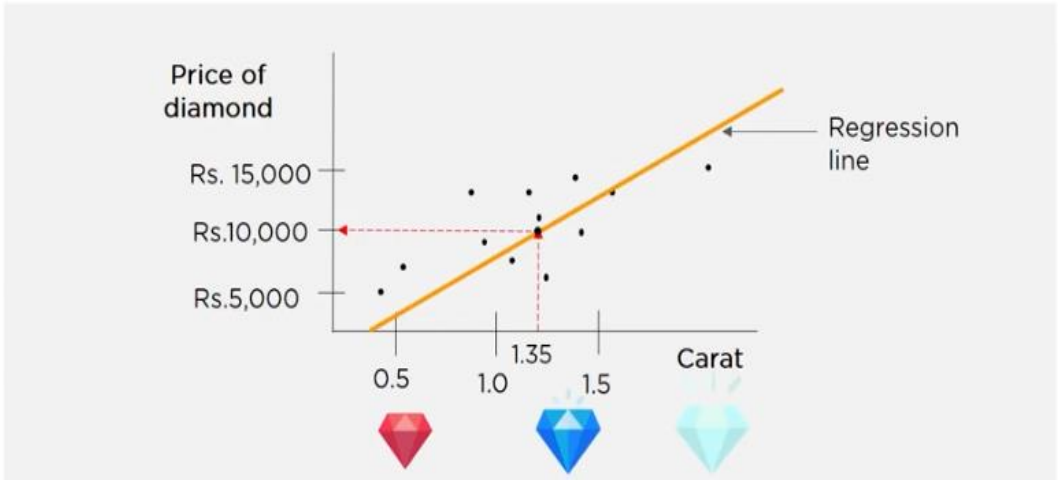
After the regression line is drawn, we can predict Y value for a input X value using following formula: $Y = mX + c$

m = Slope of the line
c = Y intercept

Y is dependent variable

Prediction:

Thus, using Simple Linear Regression algorithm we have implemented a successful model and predicted the price of 1.35 carat diamond to be Rs. 10,000



This model is easily built using Python packages like pandas, matplotlib, numpy

We will study this in detail in the upcoming Data Science Tutorial using Python

6. Model Evaluation

So in order to make our model more successful, we first need to calculate its current state.

There are two methods of evaluating models in data science,

1. **Hold-Out and Cross-Validation**. *The purpose of holdout evaluation is to test a model on different data than it was trained on*. *This provides an unbiased estimate of learning performance.*

2. **Cross-validation** *is a technique that involves partitioning the original observation data set into a training set, used to train the model, and an independent set used to evaluate the analysis*.

    - **To avoid over-fitting**, both methods use a test set (not seen by the model) to evaluate model performance. *If we do not obtain a satisfactory result in the evaluation, we must re-iterate the entire modelling process until the desired level of metrics is achieved.*

Common metrics used to evaluate models:

Classification metrics:

- Precision-Recall,
- ROC-AUC,
- Accuracy,
- Log-Loss

Regression metrics:

- MSPE,
- MSAE,
- R Square,
- Adjusted R Square