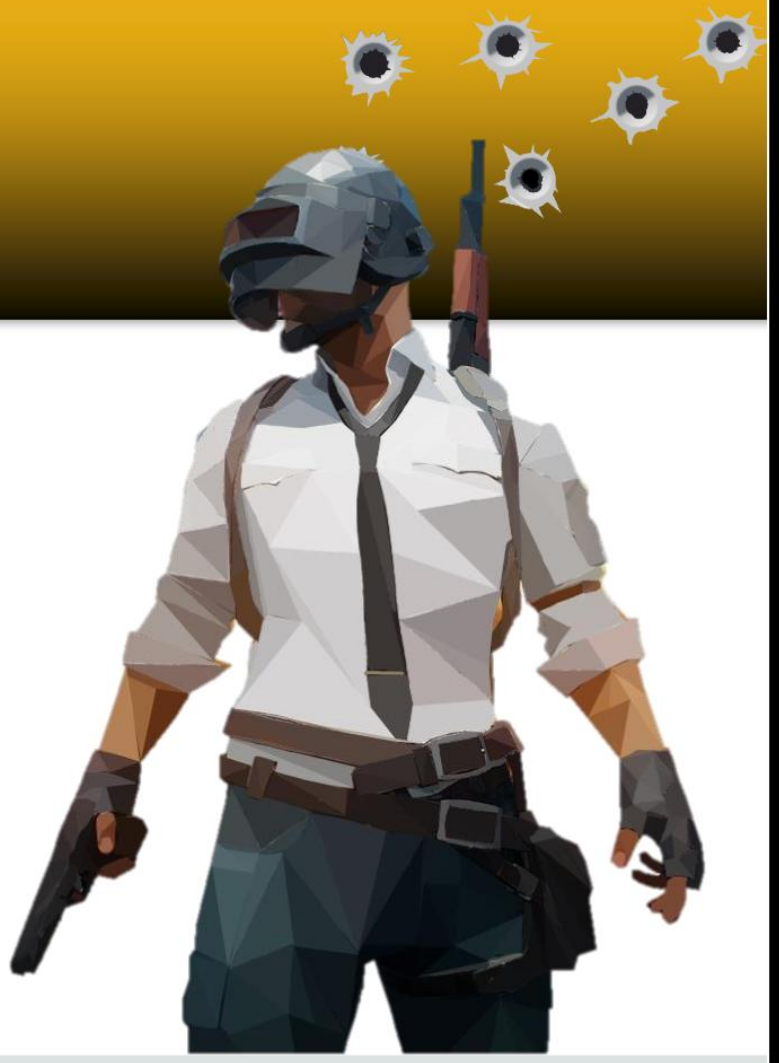

PLAYERUNKNOWN'S
BATTLEGROUNDS
A DATA SCIENCE WORKSHOP

PUBG™

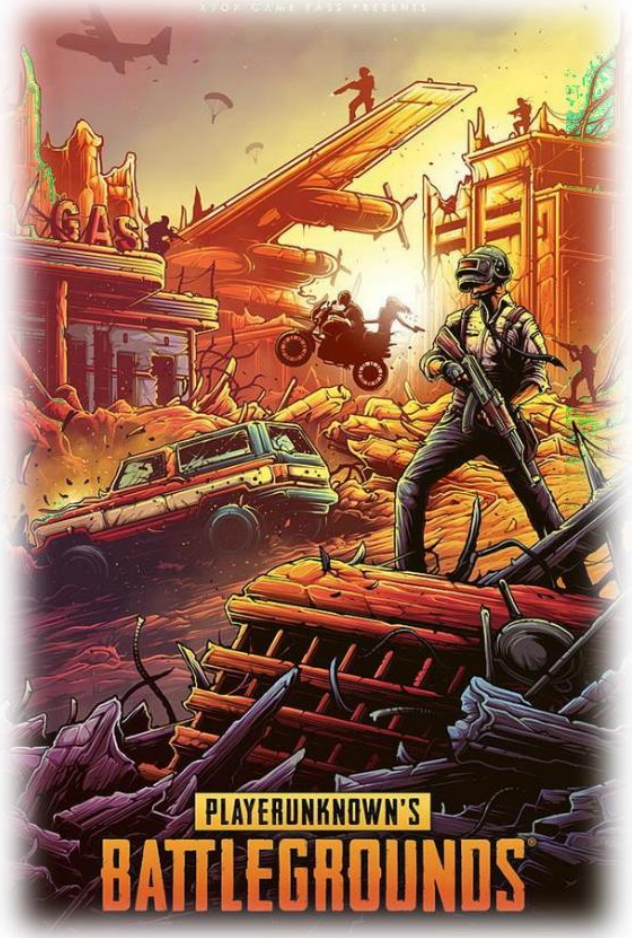


- **PlayerUnknown's Battlegrounds (PUBG)** is an online multiplayer *battle royale* game developed and published by PUBG Corporation

- **Battlegrounds** is a player versus player shooter game in which up to one hundred players fight in a *battle royale*, a type of last man standing deathmatch

Problem Statement

You are provided with a large number of anonymized **PUBG game stats**. Each row contains one player's post-game stats. Create a model which **predicts players' finishing placement** based on their final stats, on a scale from 1 (first place) to 0 (last place).



Perform the PUBG data analysis and answer the following questions:

- Does killing more people increases the chance of winning the game?
- How does total distance travelled by the player impact the winning?
- How do we catch the fraudsters in the game?
- Which features were most important while making the prediction?

MISSION OBJECTIVE

DAY 1

- Explore career opportunities in Data Science
- A day in a life of a Data Scientist
- Discuss the FAQs related to Data Scientist
- Understand how to pre-process the data
- Explore the data to generate meaningful insights
- Apply feature engineering to generate new features

DAY 2

Career Opportunities in Data Science

Skills/ Roles	Data Analyst	Data Administrator	ML Engineer	BI developer	Data Engineer	Data Scientist	Data Architect
Programming	●	●	●	●	●	●	●
Business Intelligence	●	●	●	●	●	●	●
Database Management	●	●	●	●	●	●	●
Statistics	●	●	●	●	●	●	●
Data Wrangling	●	●	●	●	●	●	●
Machine Learning	●	●	●	●	●	●	●
Deep Learning	●	●	●	●	●	●	●
Mathematics & Linear Algebra	●	●	●	●	●	●	●
Cloud	●	●	●	●	●	●	●
Big Data	●	●	●	●	●	●	●

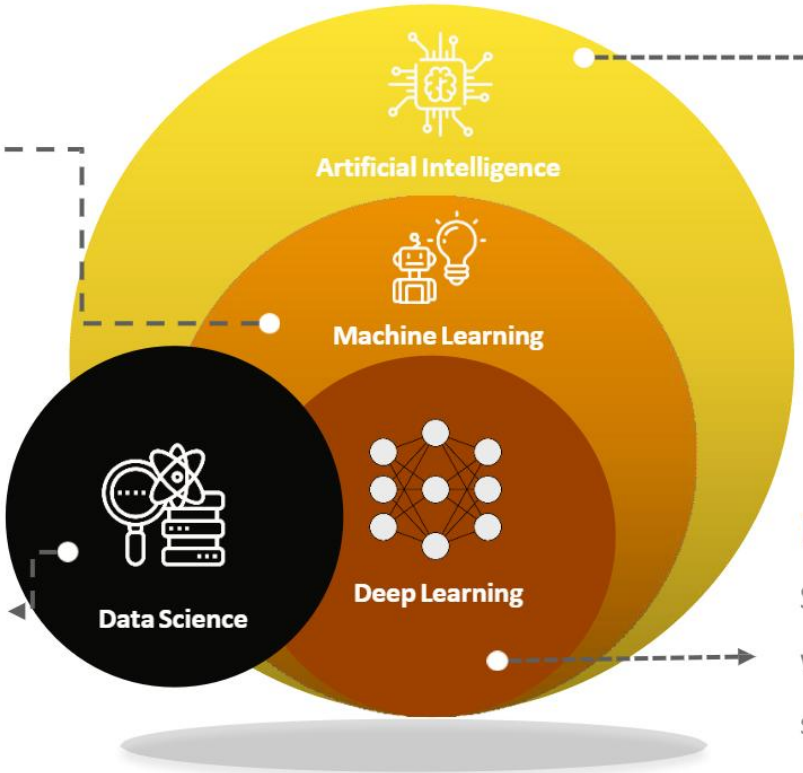
edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Myth 1: AI vs ML vs DL vs DS – Are All Same?

Machine Learning

A subset of AI which gives a machine the ability to use the stat model to learn from the data.



Artificial Intelligence

Area of computer science that emphasizes on the creation of intelligent machines that work and react like humans.

Data Science

Data science is not exactly a subset of ML, but it uses ML and DL to gain insights from both structured and unstructured data.

Deep Learning

Subset of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

Data Scientist: FAQs

WhoAm I?

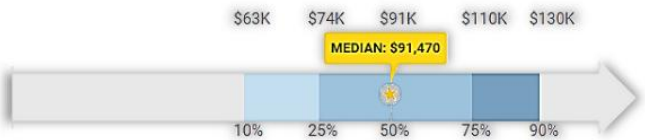
I am part analyst and a part artist. I use my technical and analytical skills to extract insights out of the data

What did I learn?

- Python/ R
- Statistical Analytics
- Predictive Modelling
- Machine Learning
- Natural Language Processing
- Deep Learning
- Data Warehousing
- Big Data
- Cloud– AWS/Azure

How much do I earn?

Data Scientist Salary (IN)



Data Scientist Salary (US)

What do I do?

- Collect data and analyse it from various angles
- Clean existing raw data and build predictive models out of it
- Identify correct business problems and give solution with visualizations

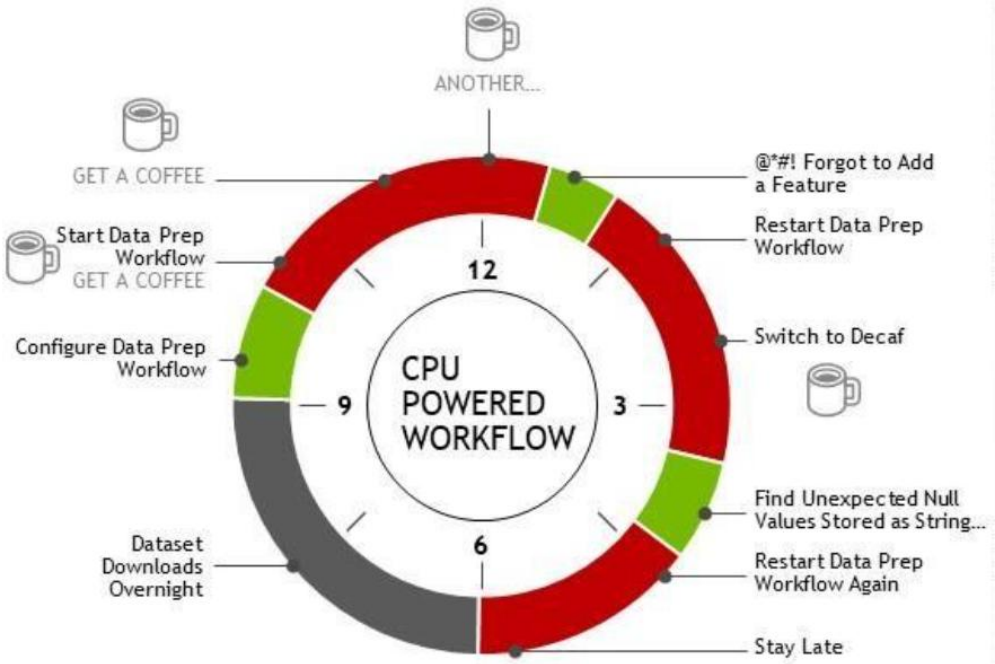
How do I help my organization?

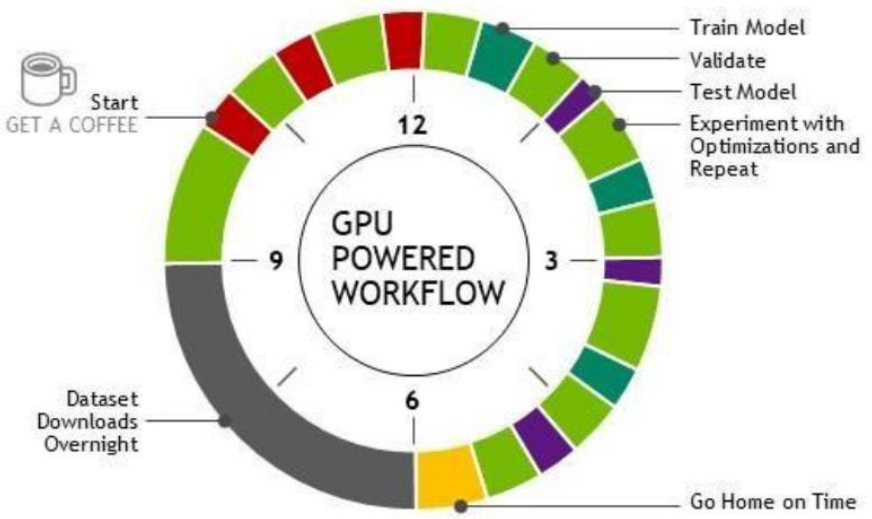
- Cost Optimization
- Develop Strategies
- Improve Operational Efficiency
- Risk Optimization
- Build Recommender System
- Increase data accuracy

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

A Day in a Life of a Data Scientist

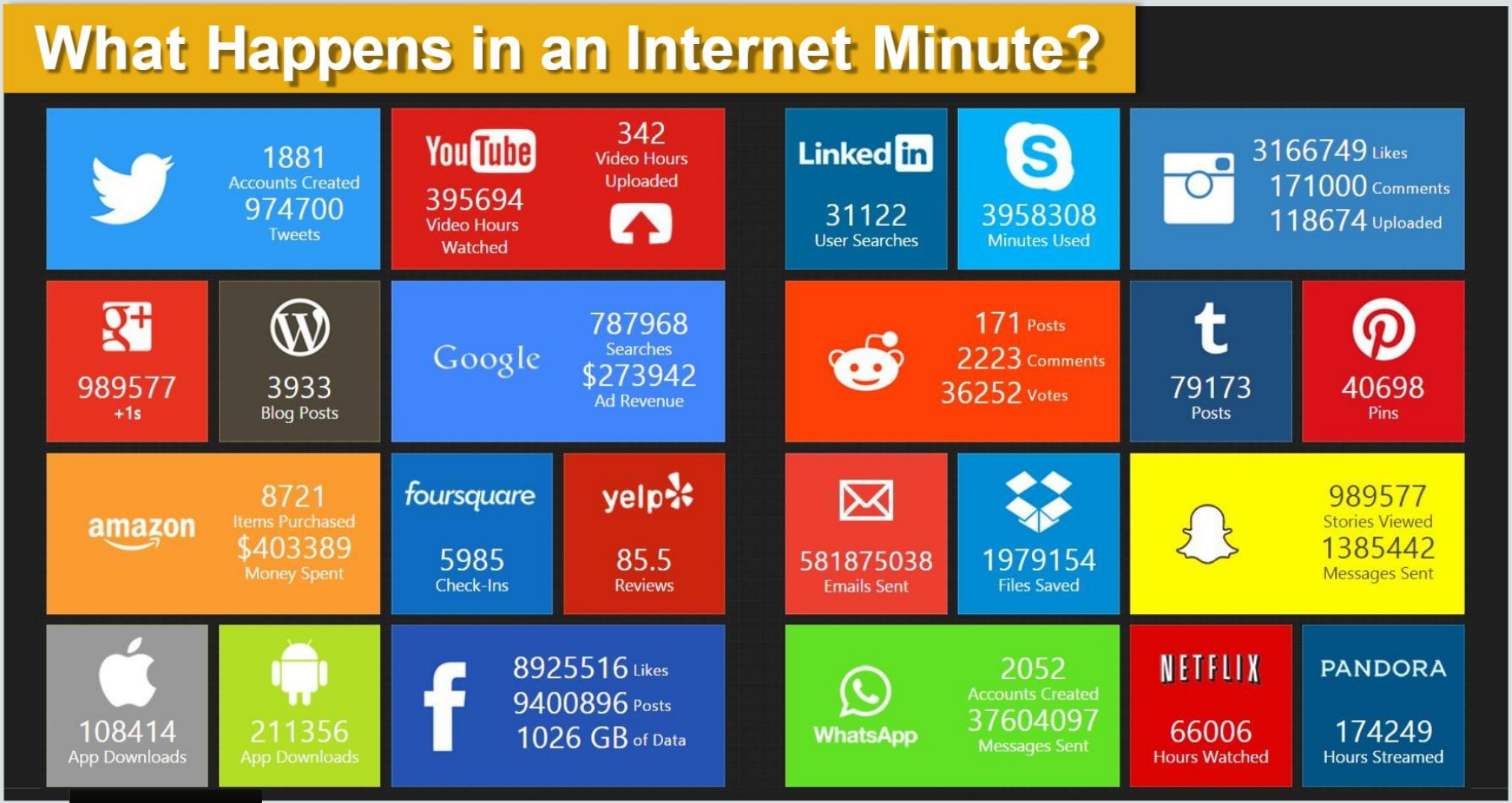




Dataset Collection | Analysis | Data Prep | Train | Inference

edureka!

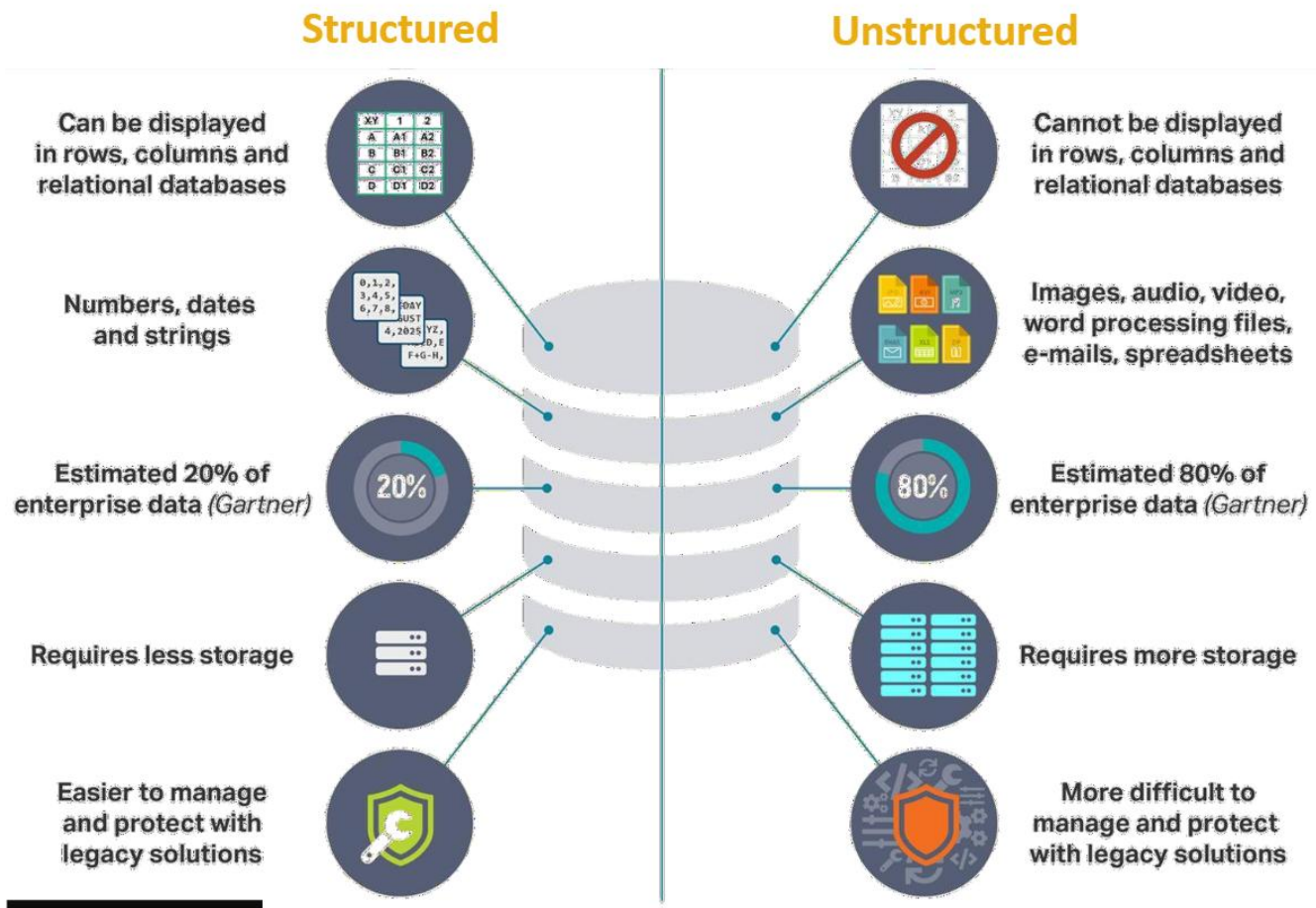
What Happens in an Internet Minute?

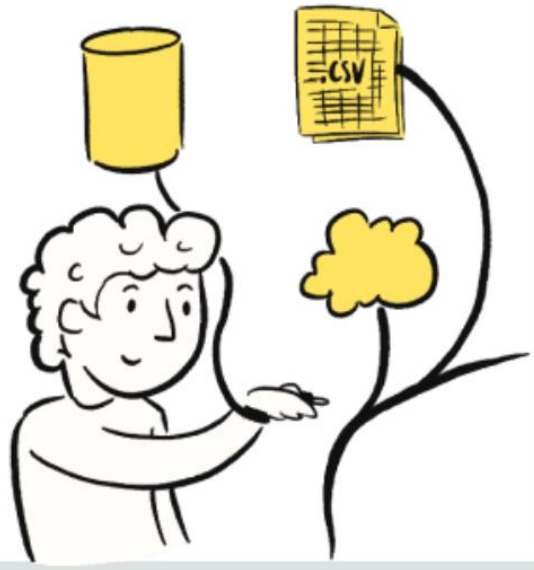


edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Varieties of Data Around us

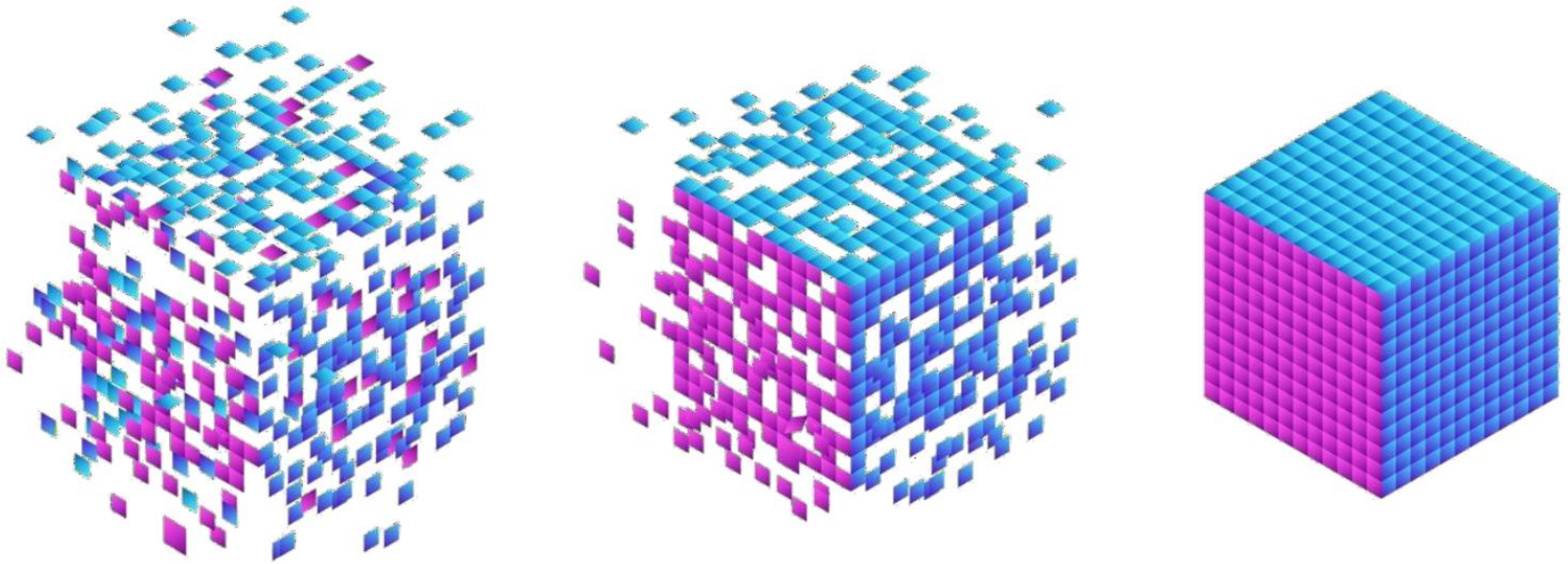




edureka!

How to Make Sense of Data?

DATA ▶ ANALYTICS ▶ **DECISIONS**



edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

How to Make Sense of Data?

DATA

- Data Collection
- **Data Preprocessing**
- Data Transformation
- **Feature Engineering**
- Feature Scaling

ANALYTICS

- Descriptive Analysis
- Predictive Analysis
- **Exploratory Data Analysis (EDA)**
- Probabilistic Theory
- Inferential Statistics

DECISIONS

- **Does killing more people increases the chance of winning the game?**
- **Which features were most important while making the prediction?**

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

DATA PREP

edureka!

Why do we Need Data Pre-processing?

- Scrapped data from real world is dirty!
- **Incomplete:** missing attribute values, contains only aggregate data

Eg: sales = " "

- **Noisy:** Contains errors or outliers

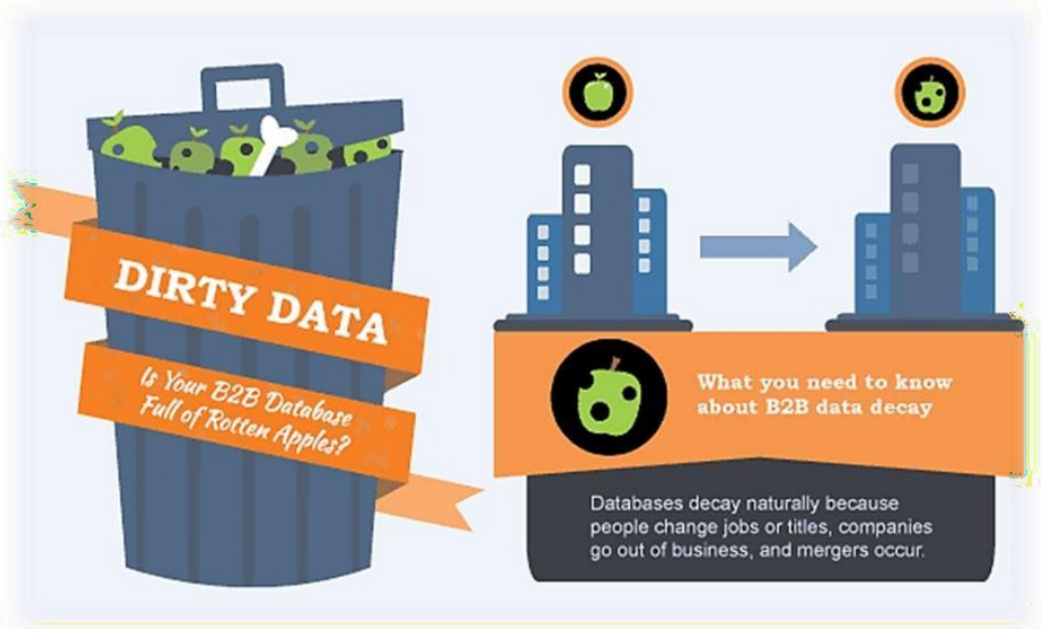
Eg: salary = "-10"

- **Inconsistent:** Contains discrepancies in code

Eg: Age = "30", Birthday ="10th August, 93", "10/08/1993

Eg: Was rating: "1,2,3", now rating: "A, B, C"

Eg: Duplicate records



Data Pre-processing Techniques

- **Step 1:** Import the libraries and dataset
- **Step 2:** Take care of Missing **Data** in Dataset
- **Step 3:** Convert numbers stored as string into numbers
- **Step 4:** Find the outliers and remove the noise
- **Step 5:** Remove duplicate values
- **Step 6:** Remove attributes with low co-relation with the target variable
- **Step 7:** Use dimensionality reduction to reduce the data
- **Step 8:** Encode the categorical **data**
- **Step 9:** Split the data into Train and Test Set
- **Step 10:** Use Feature Scaling to standardize the independent features

edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Myth 2: Its All About the Right Algorithm!

"I can't find an efficient algorithm, but neither can all these famous people."

No, its all about the data!

For better results, a model needs **quality** data (right data) in **quantity** (more data). Also building a good model requires the access to the best data from real-world scenarios.

Copyright © edureka and/or its affiliates. All rights reserved.

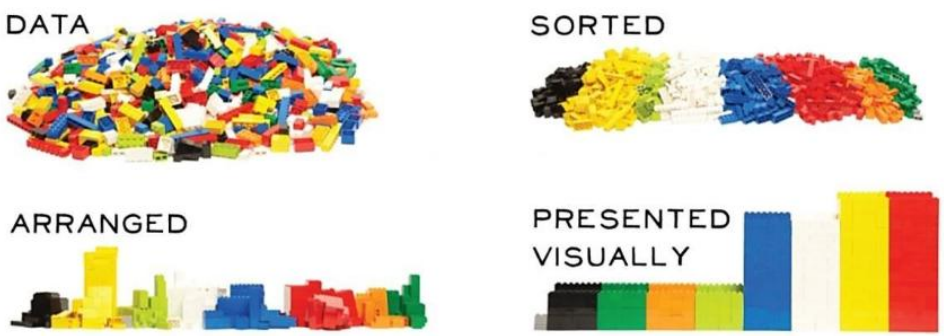
EXPLORATORY

edureka!

Exploratory Data Analysis (EDA)

Analysing the data to discover patterns, anomalies, or test hypothesis and assumptions with summary statistics and graphical representations

- **Data:** All the main content
- **Sorted:** Understand relationships and every small details
- **Arranged:** Have a distribution plan, be creative and narrative
- **Presented Visually:** Present the data visually and generate meaningful insights



edureka!

Why EDA?

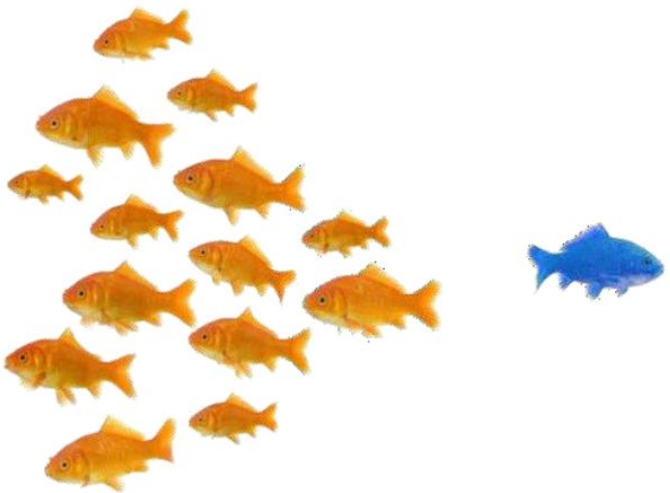
- Maximize insight into a data set
- Uncover underlying structure
- Extract important variables
- Detect outliers and anomalies
- Test underlying assumptions

Copyright © edureka and/or its affiliates. All rights reserved.

EDA: Outlier Detection

A value that "lies outside" most of the other values in a set of data.

For example: In the scores 30, 38, 4, 32, 185, 33, 36, 40 both **4** and **185** are "**outliers**".

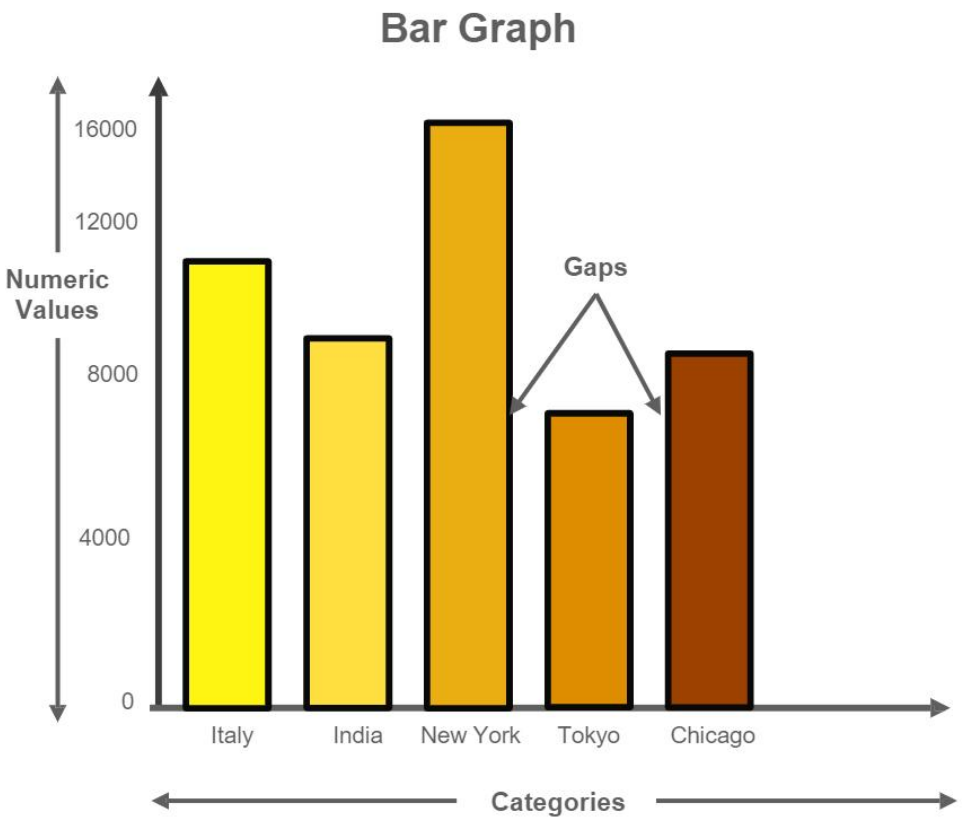


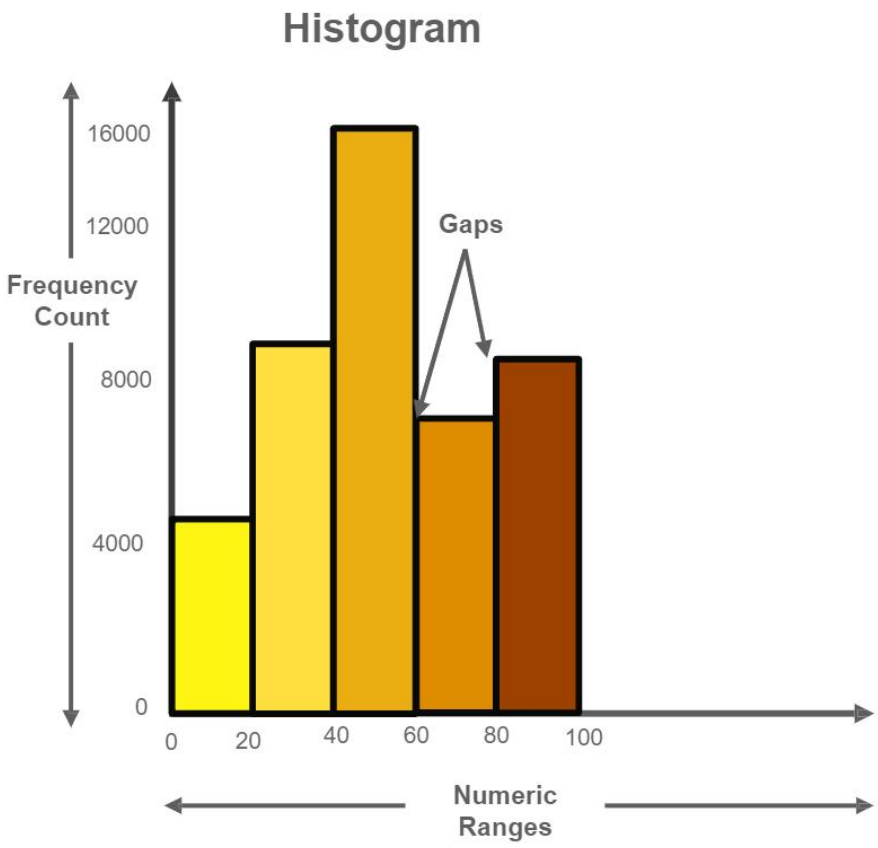
NOTE Removing outliers from the dataset helps improving the model

Copyright © edureka and/or its affiliates. All rights reserved.

EDA: Graphs – Bar Graph vs Histogram

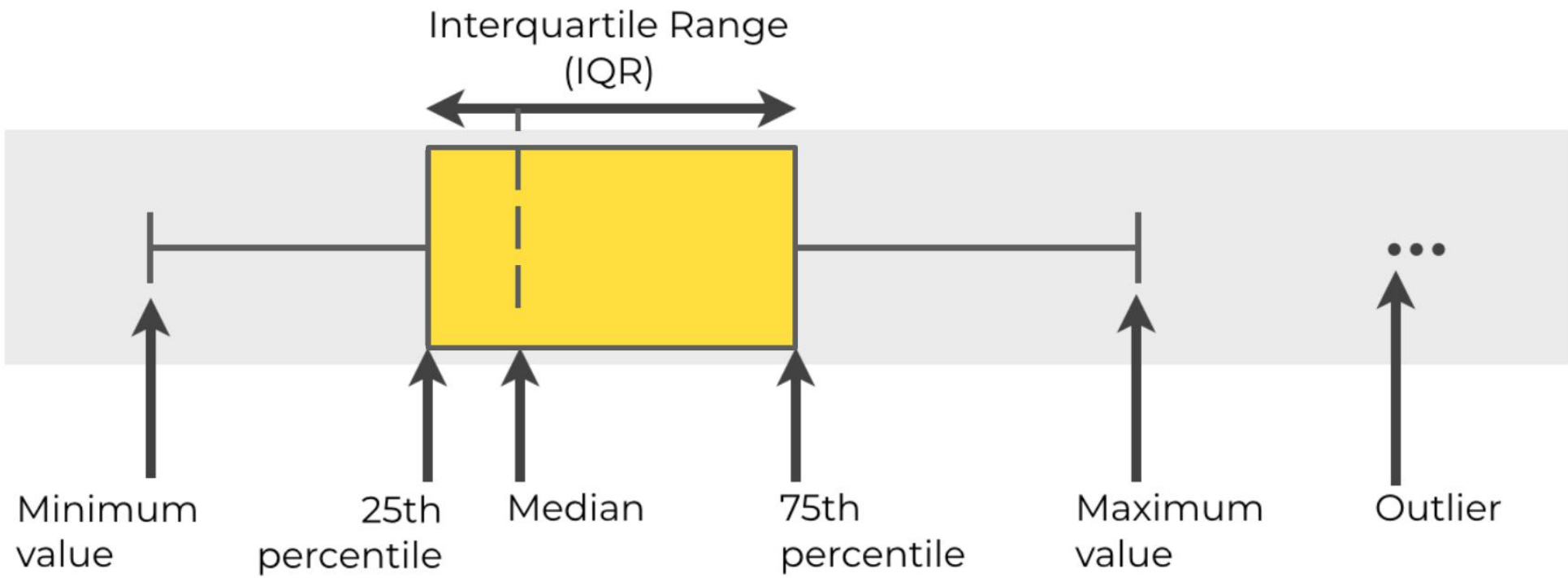
EDA: Graphs – Bar Graph vs Histogram



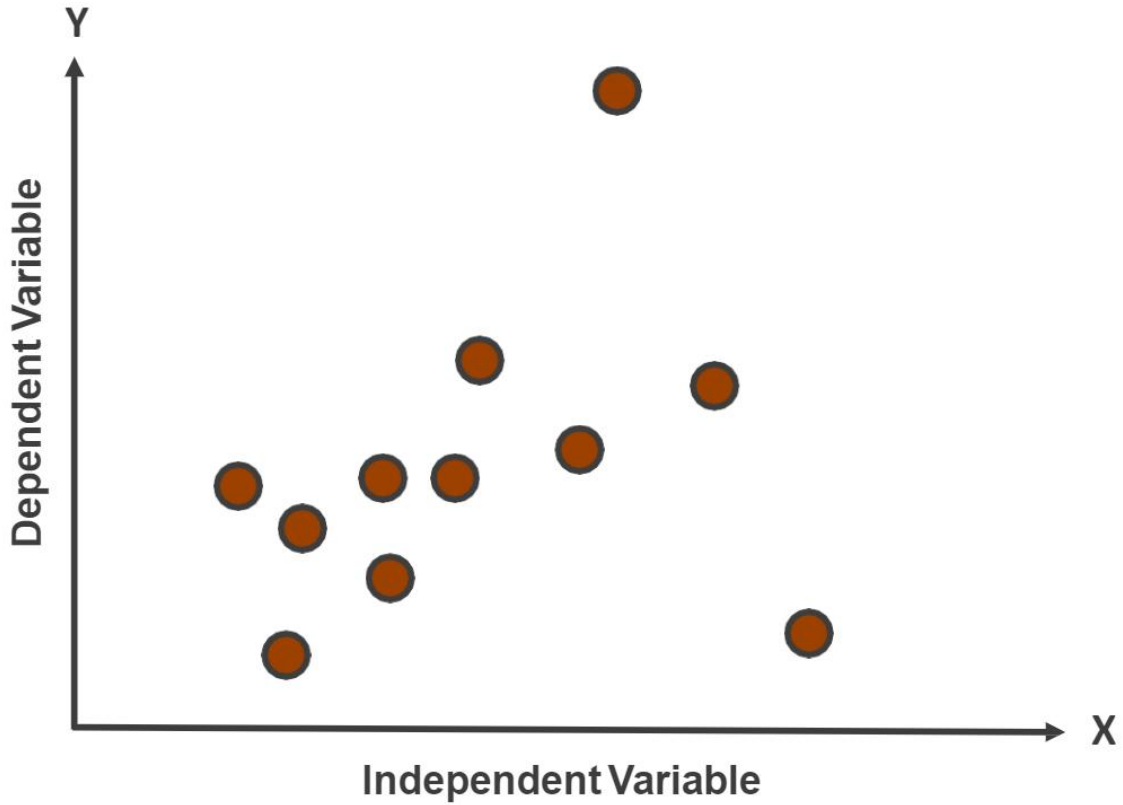


edureka!

EDA: Graphs – Box Plot

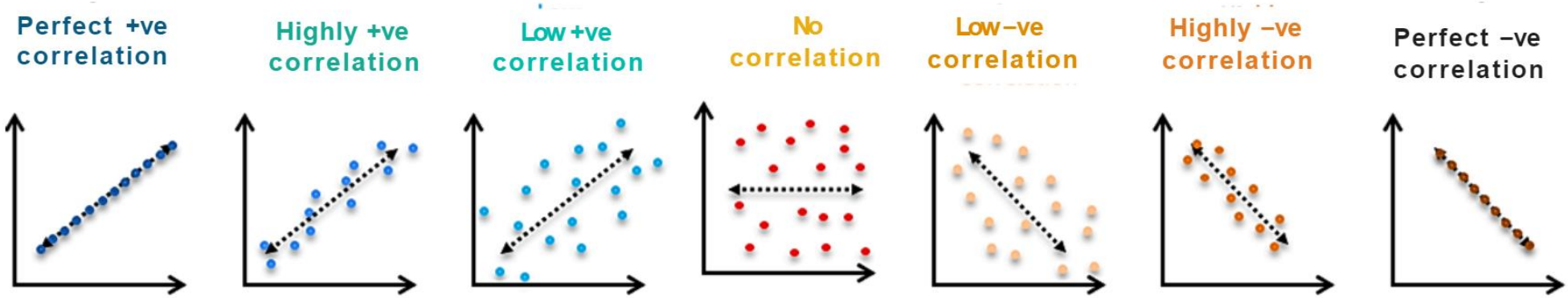


EDA: Graphs – Scatter Plot



edureka!

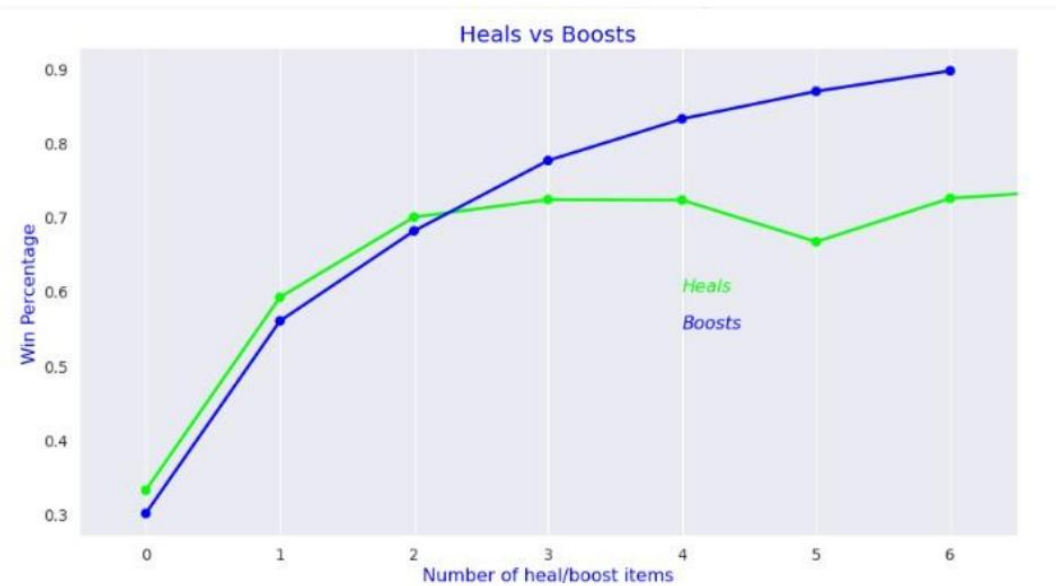
EDA: Graphs – Scatter Plot - Correlation



edureka!

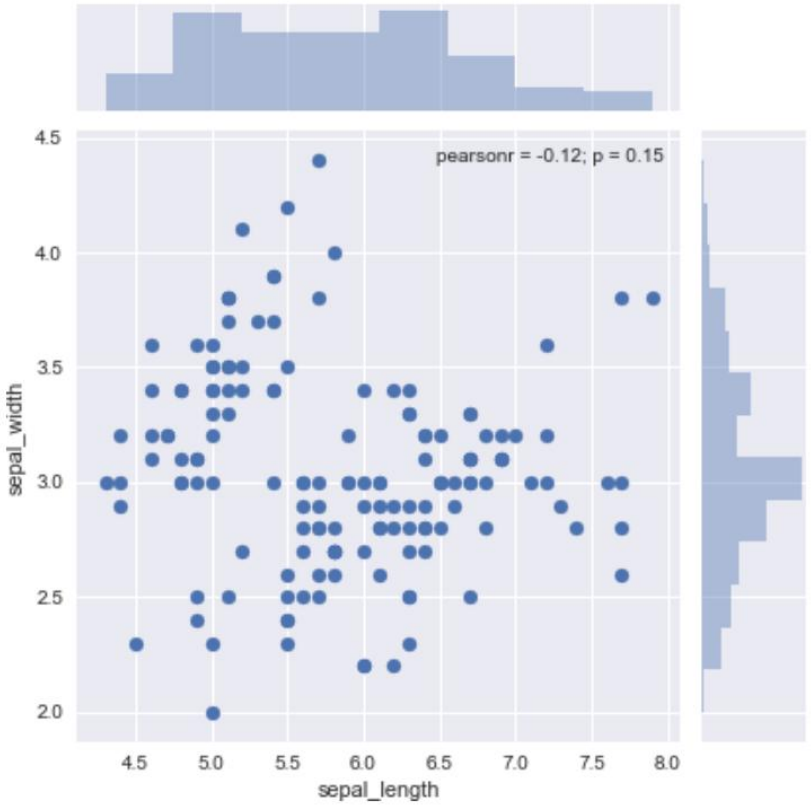
EDA: Graphs – Point Plot

Point plots is used for comparisons between different levels of one or more categorical variables



EDA: Graphs – Joint Plot

Used to quickly visualize and analyse the relationship between two variables and describe their individual distributions on the same plot.

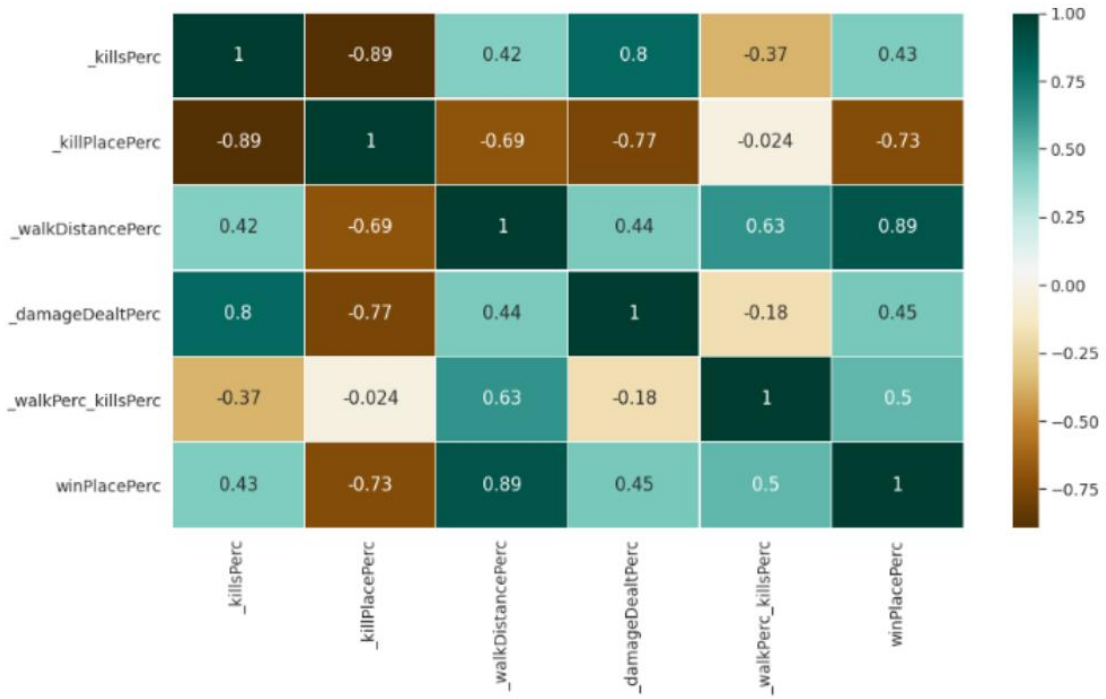


edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

EDA: Graphs – Heat Map - Pearson Correlation

- Each cell in the table shows the correlation between two variables
- Closer the value is to 1, higher it is correlated and vice versa
- Value denotes the strength of correlation
- The sign denotes whether the variables are positively or negatively correlated



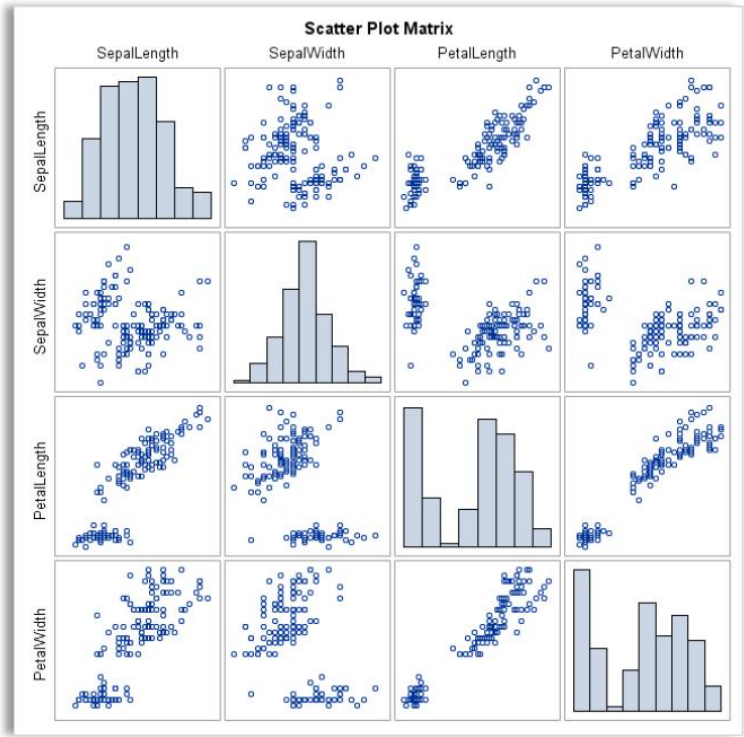
edureka!

Copyright © edureka and/or its affiliates. All rights reserved.

Q: Closely analyse the correlation plot on the right and answer the questions that follows:

1: Which of the variables are having a high correlation?

2. Why do we have a histogram all along the diagonal?



QUIZ time

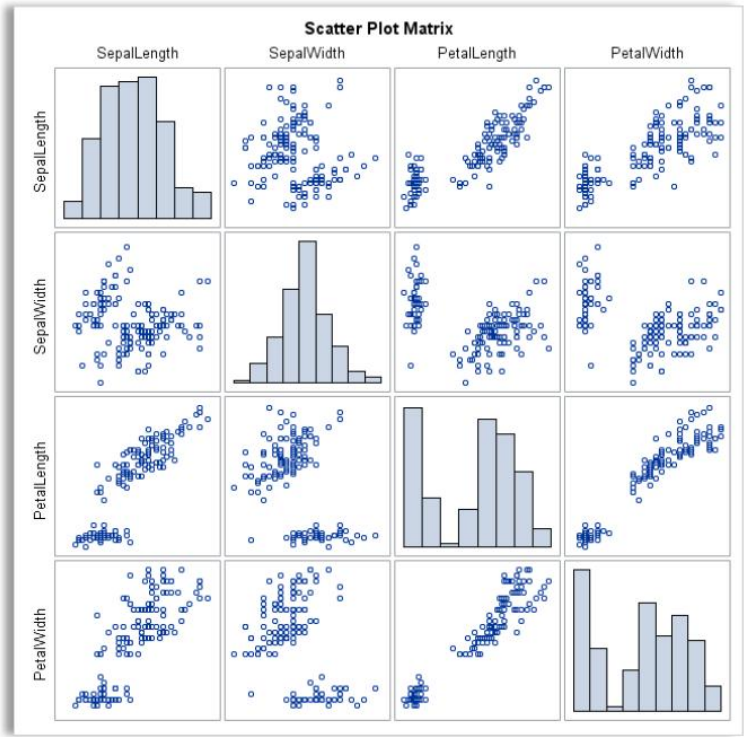
Q: Closely analyse the correlation plot on the right and answer the questions that follows:

1: Which of the variables are having a high correlation?

PetalWidth – PetalLength, SepalLength-PetalLength, SepalLength-PetalWidth

2. Why do we have a histogram all along the diagonal?

Diagonal elements have correlation 1, and since you are plotting same variable vs same variable so you are getting a frequency count of it.



edureka!

FEATURE E

edureka!

Feature Engineering

Feature engineering is the process by which existing attributes of data is used to create new attributes (**features**), that can be used to train a predictive model.

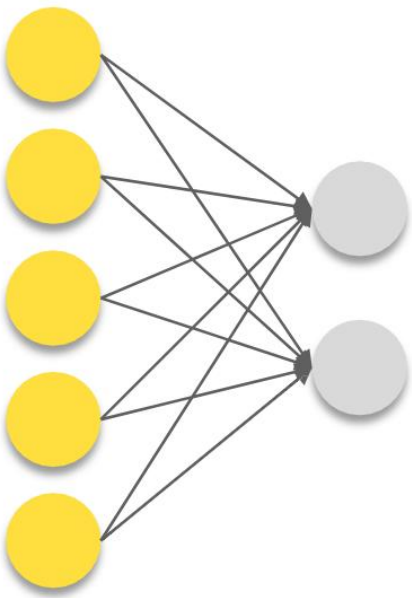
Item_ID	Item_Weight	Item_Price	Price_per_Weight
FDA15	9.3	249.81	26.86
DRC01	5.9	48.27	8.15
FDN15	17.5	141.62	8.09
FDX07	19.2	182.10	9.48

Existing Features

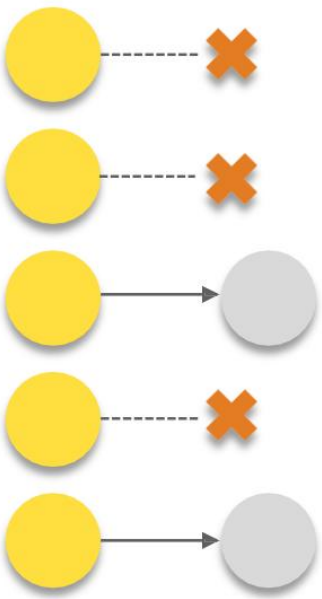
New Features

Feature Engineering: Types

Feature Extraction

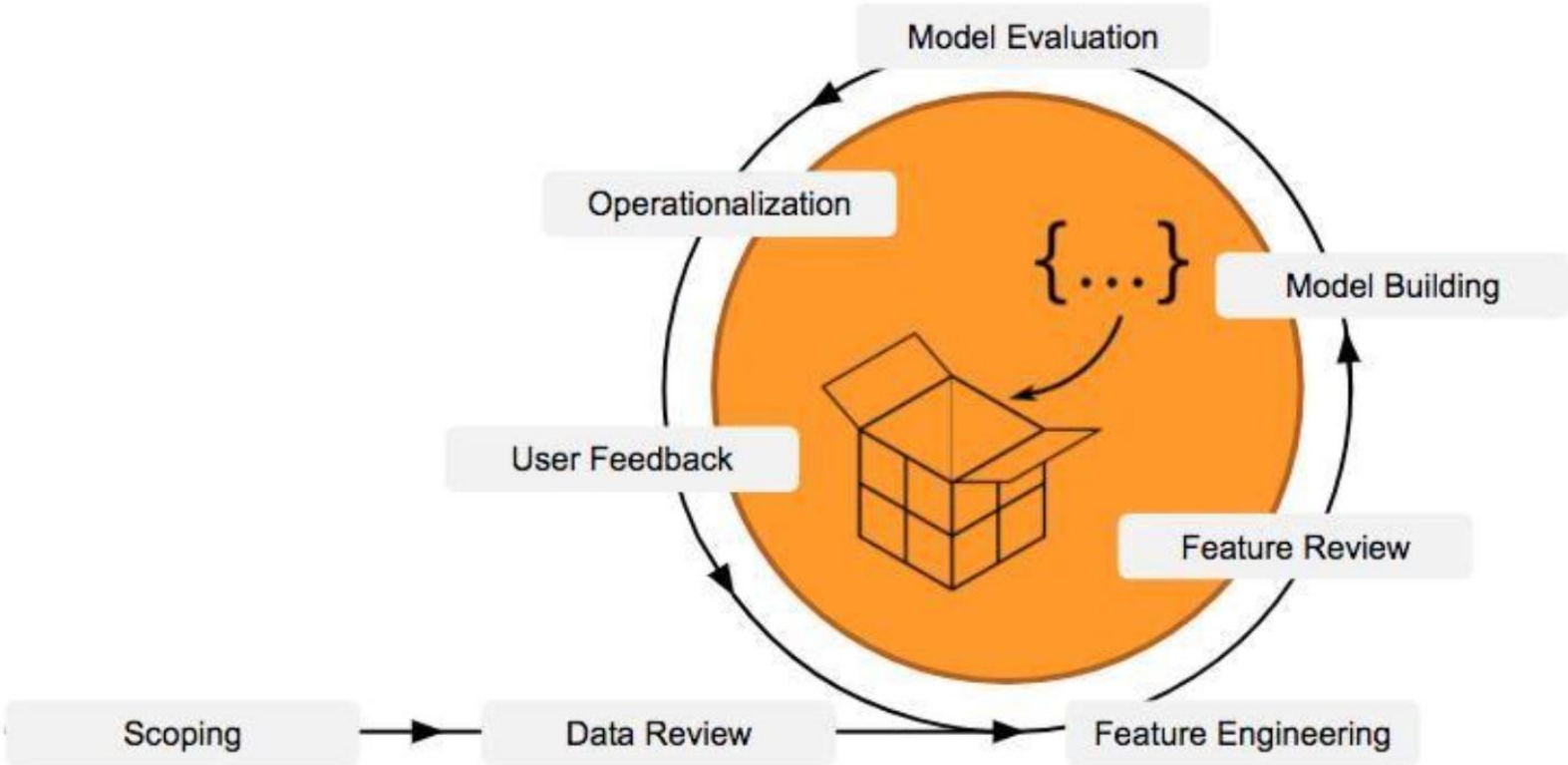


Feature Selection



edureka!

Feature Engineering: Cycle



DATA SCIENCE @edureka!

edureka!

Courses	Python for Data Science	Data Science Masters	Data Science PGP by IIT Guwahati
Python Basics	✓	✓	✓
Machine Learning	✓	✓	✓
Statistics		✓	✓
Deep Learning		✓	✓
Natural Language Processing			✓
Sequence Learning			✓
Reinforcement Learning			✓
Apache Spark and Scala		✓	
Tableau		✓	
Data Science using R		✓	

Features	Python for Data Science	Data Science Masters	AIML PGD NIT Warangal
Instructor-led LIVE Session	✓	✓	✓
24/7 Doubt Clearing Support	✓	✓	✓
Capstone Project	✓	✓	✓
Cloud Lab	✓	✓	✓
Valid Certification	✓	✓	✓
Career Assistance			✓
Alumni Status			✓
Number of Projects/ Assignments /Case Studies/ Hands-on	15+	50+	120+
Hours of LIVE Class	42 hours	250+ hours	420+ hours
Price	₹21,995	₹ 89,999	₹2,22,450

edureka!

Thank You

For more information please visit our website
www.edureka.co