Understanding Structured Data in Machine Learning
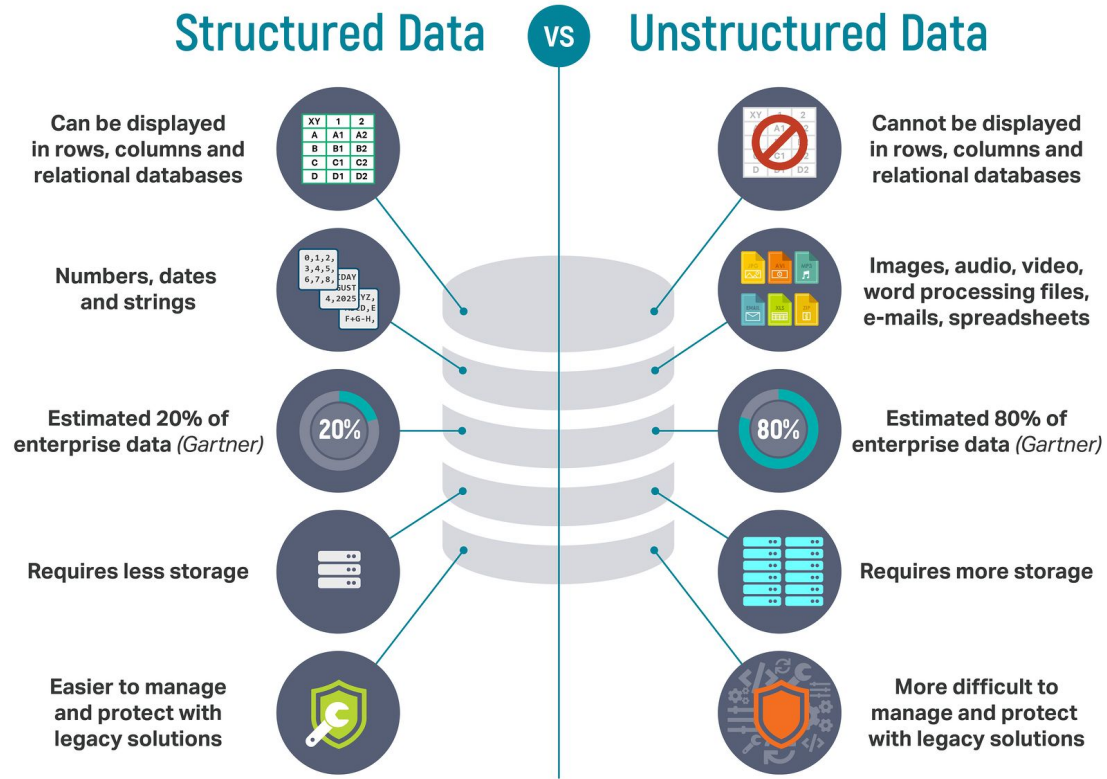

Aditya Sharma

Junior Data Scientist
Tata Consultancy Services, Ltd.

Topics covered

- Types of Data

- Exploratory Data Analysis

- Data Wrangling

Types of Data

Structured Data vs Unstructured Data

Structured Data	Unstructured Data
Can be displayed in rows, columns and relational databases	Cannot be displayed in rows, columns and relational databases
Numbers, dates and strings	Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 20% of enterprise data *(Gartner)*	Estimated 80% of enterprise data *(Gartner)*
Requires less storage	Requires more storage
Easier to manage and protect with legacy solutions	More difficult to manage and protect with legacy solutions

Need for EDA *(Exploratory Data Analysis)*

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to :
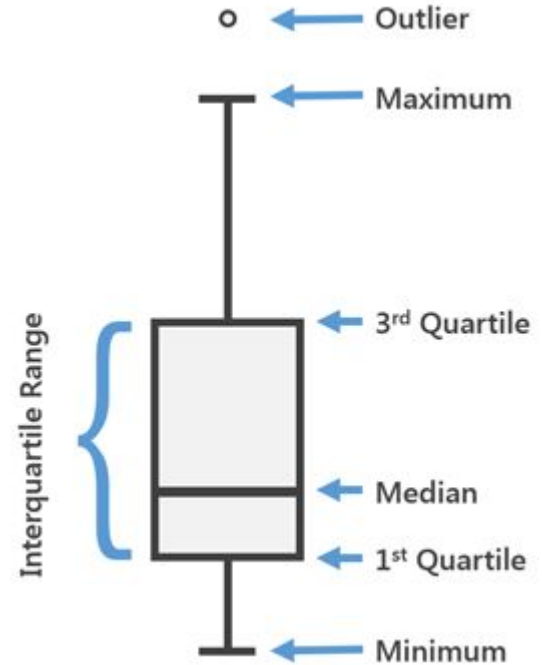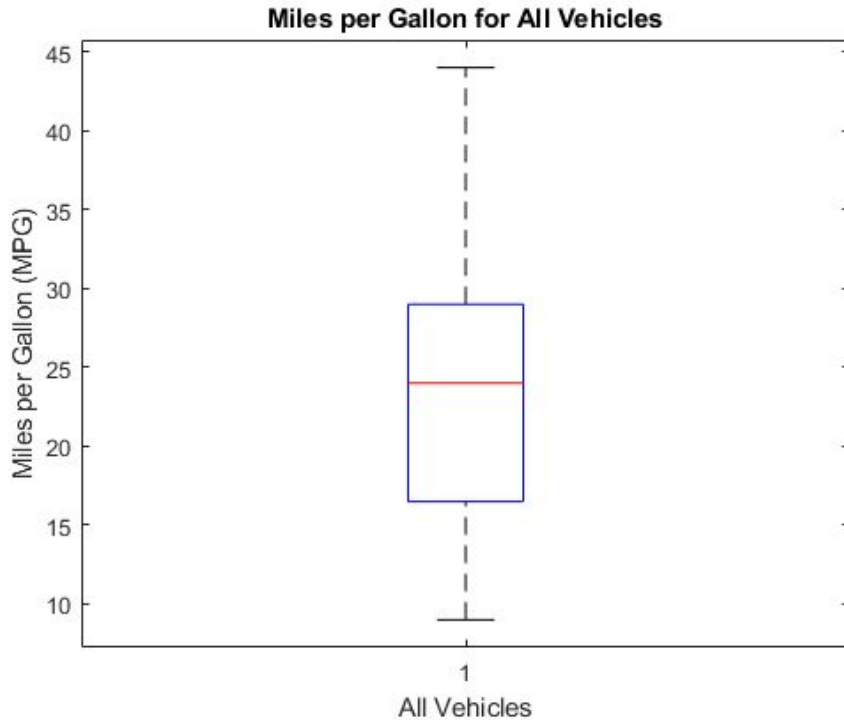➔ Discover patterns
➔ Skewness
➔ Correlation
➔ Spot anomalies
➔ Test hypothesis
➔ Check assumptions with the help of summary statistics and graphical representations.
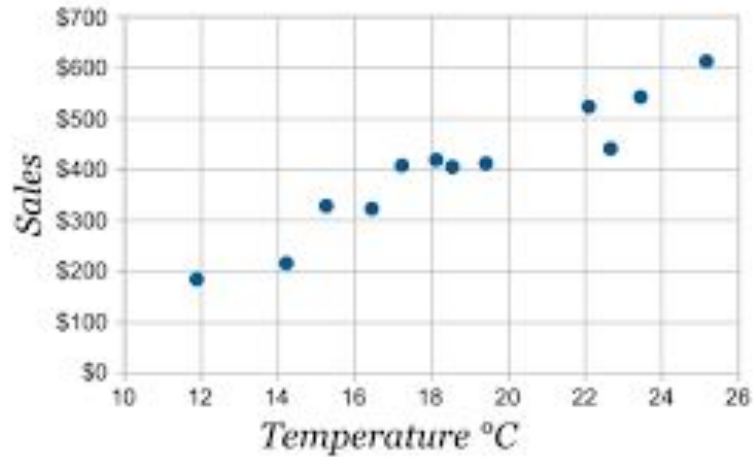
Some common visualizations

- Box plot
- Violin plot & Swarm plot
- Pie plot

- **Histograms**
- Dist plot

- Point plot
- Categorical data plots
- **Correlation matrix**

- Scatter plot
- **Pair plot**

Know more at **datavizproject.com**

Box plot (Univariate analysis)

Miles per Gallon for All Vehicles

Scatter plot (Bivarate Analysis)

Data Wrangling (Cleaning)

- Feature selection
- Delete/Impute missing values
- Convert to categorical/one hot encoded
- Scaling

Thank you!

linkedin.com/in/aditya1601