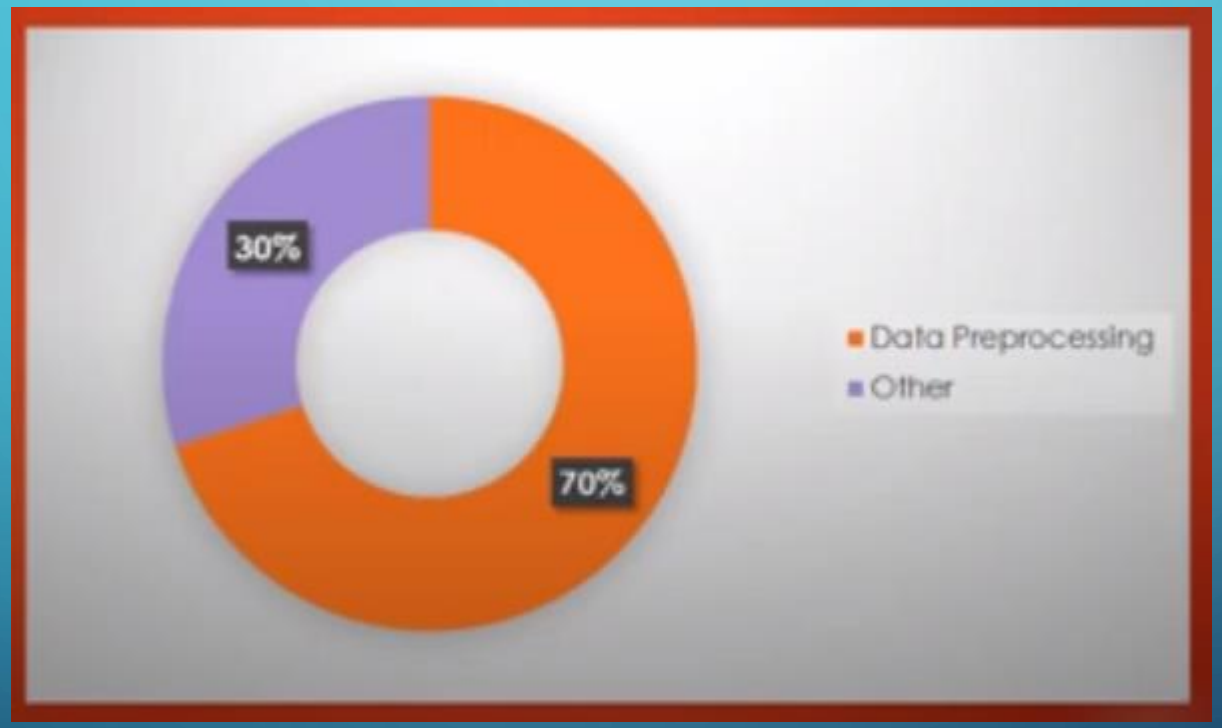
# Data Preprocessing

## Data Scientist most Spending Time!




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

# Data Preprocessing

# Data Preprocessing


To Convert raw data into valuable data,
we use Data Preprocessing.

# Data Preprocessing

## Why we use data preprocessing?

To perform better future prediction results, we have need to process data.

# Data Preprocessing

**Data Preprocessing Techniques**

- ✓ Data Cleaning
- ✓ Data Integration
- ✓ Data Reduction
- ✓ Data Transformation

# Data Preprocessing Techniques

## 1. Data Cleaning

Data cleaning is the process of fixing or removing incorrect, incomplete, corrupted, duplicate data within a dataset.

- ✓ Removing Outlier
  - Min Outlier
  - Max Outlier

- ✓ Handling Missing Values (Nan Values)

| Age | Income |
|---|---|
| 32 | 60,000 |
| 33 | Nan |
| 36 | 90,000 |

| item | Price | |
|---|---|---|
| 1 | 60 | Min Outlier |
| 2 | 650 | Valuable Data |
| 3 | 900 | |
| 4 | 950 | |
| 5 | 550 | |
| 6 | 56,890 | Max Outlier |

- ✓ Removing Strings
  - Removing Strings with Integer
  - Find and Remove Strings in the Dataset

| Age | Income |
|---|---|
| 32 | 60,000as |
| 33 | 65,00as |
| 36 | 90,000as |

# Data Preprocessing Techniques

## 2. Data Integration

Combining the data from different source
is called Data Integration.

Google Analytics

Join

Data Integration

# Data Preprocessing Techniques

## 2. Data Integration

Combining the data from different source is called Data Integration.

| Age | Class |
|---|---|
| 12 | 4 |
| 15 | 8 |

| City | Area |
|---|---|
| Lahore | Gulberb |
| Islamabad | Askari-10 |

**After Integration**

| Age | Class | City | Area |
|---|---|---|---|
| 12 | 4 | Lahore | Gulberb |
| 15 | 8 | Islamabad | Askari-10 |

# Data Preprocessing Techniques

## 3. Data Reduction

Reducing the data volume to fasten the model training.

# Data Preprocessing Techniques

## 3. Data Reduction

Reducing the data volume to fasten the model training.

| Mix | MN | Into | As |
|---|---|---|---|
| 45 | bdc | 146 | Gge5 |
| 33 | bd33 | 14 | 55gh |
| 789 | x2d | 365 | c25c |

| Mix | Into |
|---|---|
| 45 | 146 |
| 33 | 14 |
| 789 | 365 |

**Data Reduction Techniques**

- ✓ Dimension Reduction
- ✓ Numerosity Reduction

# Data Preprocessing Techniques

## 4. Data Transformation

Data Transformation is the last stage of Data Preprocessing.
After this our data is ready to fit on the model.

Techniques of Data Transformation

**Feature Engineering (Feature Scaling)**

- ✓ Min Max Scaler
- ✓ Standard Scaler
- ✓ Max Abs Scaler
- ✓ Robust Scaler
- ✓ Quantile Transformer Scaler
- ✓ Log Transformation

etc....