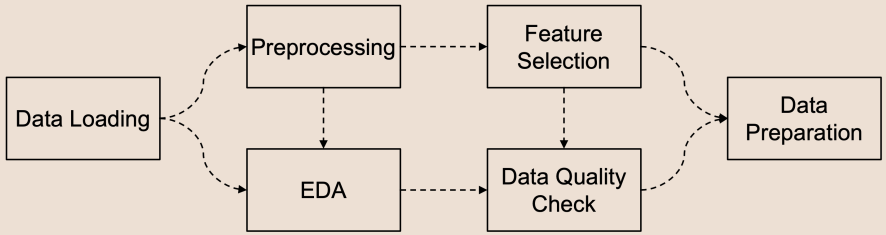
# Data Pipeline with Modeva :: CHEATSHEET

Modeva-AI

**Data Pipeline** is designed to streamline end-to-end data workflows such as data loading, preprocessing, feature selection, EDA, data quality check and preparing data for modeling.




**INSTALLATION:** Modeva is a Python package you can install directly by pip.

```
pip install modeva
```

**DataSet Class:**

```
from modeva import DataSet
ds = DataSet()
```

## Data Loading

**Load built-in data**

```
ds.load("CaliforniaHousing")
ds.load("BikeSharing")
ds.load("SimuCredit")
ds.load("TaiwanCredit")
```

**Load user data**

```
ds.load_dataframe()
ds.load_csv()
ds.load_spark()
```

## Preprocessing

**Define preprocessing steps**

```
ds.reset_preprocess()
ds.impute_missing()
ds.scale_numerical()
ds.bin_numerical()
ds.encode_categorical()
```

**Execute preprocessing**

```
ds.preprocess()
```

## EDA (Exploratory Data Analysis)

**Data exploration**

```
ds.eda_1d()
ds.eda_2d()
ds.eda_3d()
ds.eda_correlation()
```

**Dimension reduction**

```
ds.eda_pca()
ds.eda_umap()
```

## Data Quality Check

```
ds.summary()
ds.data_drift_test()
ds.subsample_random()
ds.set_active_samples()
ds.set_inactive_samples()
```

**Outlier detection**

```
ds.detect_outlier_pca()
ds.detect_outlier_cblof()
ds.detect_outlier_isolation()
```

## Feature Selection

**Feature selection**

```
ds.feature_select_corr()
ds.feature_select_xgbpfi()
ds.feature_select_rcit()
```

**Set active/inactive features**

```
ds.set_active_features()
ds.set_inactive_features()
```

## Data Preparation

**Target and task setup**

```
ds.set_target()
ds.set_task_type()
ds.set_sample_weight()
ds.set_feature_type()
```

**Raw data split**

```
ds.set_random_split()
ds.set_train_idx()
ds.set_test_idx()
ds.set_raw_extra_data()
```