

MSc - Data Mining

Topic 03 : Exploratory Data Analysis

Part 01 : Exploratory Data Analysis

Dr Bernard Butler and Dr Kieran Murphy

Department of Computing and Mathematics, WIT.
(bbutler@tssg.org and kmurphy@wit.ie)

Spring Semester, 2021

Outline

- EDA Process
- Datasets = Tips, Titanic and Algae Blooms
- Identifying and resolving issues (missing value, outliers)
- Generating ToDo list for Feature Engineering/Transformation/Selection

Data Mining

Introduction

Foundation

Data Handling

Exploratory Data
Analysis

Data Modelling
Fundamentals

Data Modelling
Advanced

Rule Based

Association Rules

Recommender Systems

Unsupervised

Clustering

Anomaly Detection

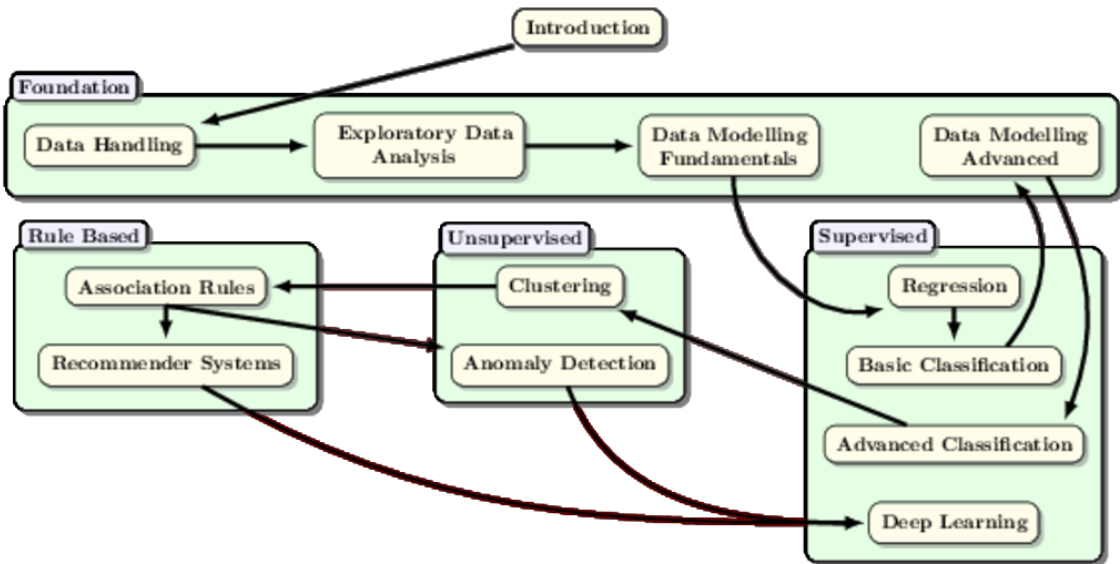
Supervised

Regression

Basic Classification

Advanced Classification

Deep Learning



1. First Pass — Load Dataset and Initial Clean

First Pass — Load Dataset and Initial Clean

- Load dataset
- Check variables names
- Verify variable types
- Identify (and possibly address) missing values

Tips — Load

```
df = pd.read_csv("tips.csv")
print(df.shape)
df.head(10)
```

(244, 7)

| | total_bill | tip | sex | smoker | day | time | size |
|---|------------|------|--------|--------|-----|--------|------|
| 0 | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 |
| 1 | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 |
| 2 | 21.01 | 3.50 | Male | No | Sun | Dinner | 3 |
| 3 | 23.68 | 3.31 | Male | No | Sun | Dinner | 2 |
| 4 | 24.59 | 3.61 | Female | No | Sun | Dinner | 4 |
| 5 | 25.29 | 4.71 | Male | No | Sun | Dinner | 4 |
| 6 | 8.77 | 2.00 | Male | No | Sun | Dinner | 2 |
| 7 | 26.88 | 3.12 | Male | No | Sun | Dinner | 4 |
| 8 | 15.04 | 1.96 | Male | No | Sun | Dinner | 2 |
| 9 | 16.99 | 1.78 | Male | No | Sun | Dinner | 2 |



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
```

```
#    Column      Non-Null Count  Dtype
---
```

```
0    total_bill  244 non-null    float64
1     tip        244 non-null    float64
2     sex        244 non-null    object
3     smoker     244 non-null    object
4     day        244 non-null    object
5     time       244 non-null    object
6     size       244 non-null    int64
```

```
dtypes: float64(2), int64(1), object(4)
memory usage: 13.5+ KB
```

Tips — Load

```
df = pd.read_csv("tips.csv")
print(df.shape)
df.head(10)
```

(244, 7)

| | total_bill | tip | sex | smoker | day | time | size |
|---|------------|------|--------|--------|-----|--------|------|
| 0 | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 |
| 1 | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 |
| 2 | 21.01 | 3.50 | Male | No | Sun | Dinner | 3 |
| 3 | 23.68 | 3.31 | Male | No | Sun | Dinner | 2 |
| 4 | 24.59 | 3.61 | Female | No | Sun | Dinner | 4 |
| 5 | 25.29 | 4.71 | Male | No | Sun | Dinner | 4 |
| 6 | 8.77 | 2.00 | Male | No | Sun | Dinner | 2 |
| 7 | 26.88 | 3.12 | Male | No | Sun | Dinner | 4 |
| 8 | 15.04 | 1.96 | Male | No | Sun | Dinner | 2 |
| 9 | 14.78 | 3.23 | Male | No | Sun | Dinner | 2 |



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 244 entries, 0 to 243
```

```
Data columns (total 7 columns):
```

```
#    Column    Non-Null Count  Dtype
```

```
0    total_bill  244 non-null    float64
1     tip        244 non-null    float64
2     sex        244 non-null    object
3     smoker     244 non-null    object
4     day        244 non-null    object
5     time       244 non-null    object
6     size       244 non-null    int64
```

```
dtypes: float64(2), int64(1), object(4)
```

```
memory usage: 13.5+ KB
```

Issue: categorical data treated as object (string).

fdd

FfdddddF

This text