ADNOC Accelerator Programme

Artificial Intelligence

Cohort 2

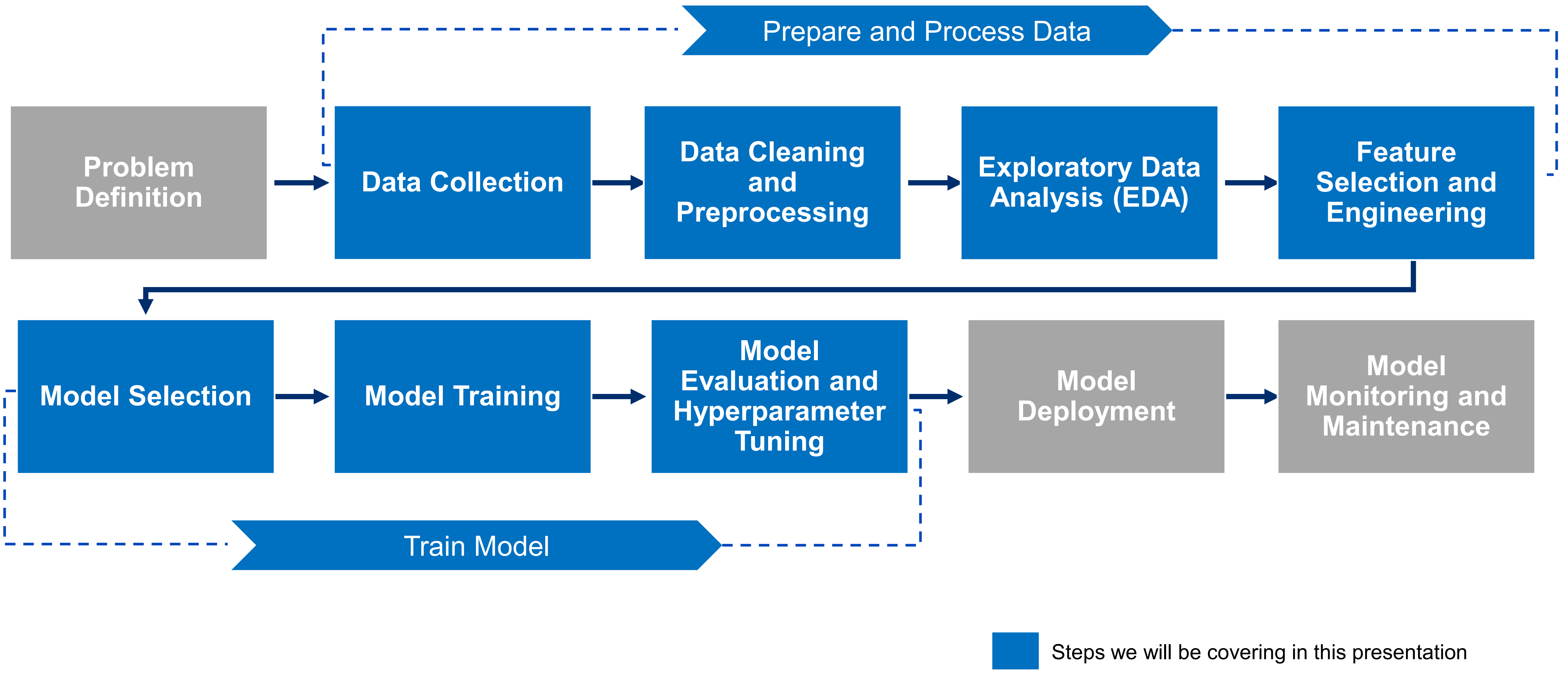
Data Transformation and Cleaning

Data Transformation and Cleaning

LEARNING OBJECTIVES

1. Utilise summary statistics for datasets
2. Prepare data spreads
3. Transform data with feature engineering

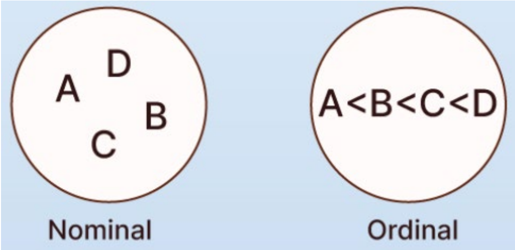
Data processing is a crucial step in the Machine Learning lifecycle



Data can be quantitative or qualitative, and have further subtypes

Data

- Qualitative (data divided into groups/labels)
  - Nominal (labels without order)
  - Ordinal (labels with order)
  - Binary
- Quantitative (data represented by numbers)
  - Continuous (measured variable; can take any value within a specified range)
  - Discrete (counter variable; can take specific value)



Datasets are explored through basic statistics and visualisation

Data columns

Pipeline Location	Pipeline Type	Cause Category	Cause Subcategory	Net Loss (Barrels)	All Costs
ONSHORE	UNDERGROUND	CORROSION	INTERNAL	0.00	5065
ONSHORE	UNDERGROUND	MATERIAL/WELD/EQUIP FAILURE	CONSTRUCTION, INSTALLATION OR FABRICATION-RELATED	0.00	216121
ONSHORE	TANK	ALL OTHER CAUSES	MISCELLANEOUS	0.00	16200
ONSHORE	ABOVEGROUND	MATERIAL/WELD/EQUIP FAILURE	MALFUNCTION OF CONTROL/RELIEF EQUIPMENT	0.00	32477
ONSHORE	UNDERGROUND	INCORRECT OPERATION	DAMAGE BY OPERATOR OR OPERATOR'S CONTRACTOR	25.00	84600
ONSHORE	UNDERGROUND	EXCAVATION DAMAGE	THIRD PARTY EXCAVATION DAMAGE	0.00	709351
ONSHORE	ABOVEGROUND	INCORRECT OPERATION	INCORRECT VALVE POSITION	0.00	16169
ONSHORE	[illegible]	CORROSION	EXTERNAL	0.00	16025
OFFSHORE	NaN	CORROSION	EXTERNAL	0.10	1450000
ONSHORE	[illegible]	MATERIAL/WELD/EQUIP FAILURE	CONSTRUCTION, INSTALLATION OR FABRICATION-RELATED	0.00	40000
ONSHORE	[illegible]	CORROSION	INTERNAL	0.00	17485
ONSHORE	[illegible]	CORROSION	INTERNAL	0.36	5040
ONSHORE	ABOVEGROUND	INCORRECT OPERATION	OVERFILL/OVERFLOW OF TANK/VESSEL/SUMP	0.00	11607
ONSHORE	ABOVEGROUND	MATERIAL/WELD/EQUIP FAILURE	MALFUNCTION OF CONTROL/RELIEF EQUIPMENT	0.35	6350
ONSHORE	ABOVEGROUND	MATERIAL/WELD/EQUIP FAILURE	PUMP OR PUMP-RELATED EQUIPMENT	0.00	3770
ONSHORE	ABOVEGROUND	OTHER OUTSIDE FORCE DAMAGE	VEHICLE NOT ENGAGED IN EXCAVATION	0.10	150001
ONSHORE	UNDERGROUND	EXCAVATION DAMAGE	OPERATOR/CONTRACTOR EXCAVATION DAMAGE	0.00	40750
ONSHORE	ABOVEGROUND	OTHER OUTSIDE FORCE DAMAGE	VEHICLE NOT ENGAGED IN EXCAVATION	976.00	776753
ONSHORE	ABOVEGROUND	MATERIAL/WELD/EQUIP FAILURE	NON-THREADED CONNECTION FAILURE	3.00	1270
ONSHORE	ABOVEGROUND	INCORRECT OPERATION	INCORRECT INSTALLATION	0.35	2034

Missing values

Categorical columns

Numerical columns

Statistics can be summarised in different ways

Count unique categories

```
df_sample.nunique()
```

```
Pipeline Location         2
Pipeline Type             4
Cause Category            7
Cause Subcategory        38
Net Loss (Barrels)      443
All Costs              2279
```


Check missing values

```
df_sample.isna().sum()
```

```
Pipeline Location      0
Pipeline Type         18
Cause Category         0
Cause Subcategory      0
Net Loss (Barrels)     0
All Costs              0
```


Show frequency of unique values

```
df_sample[["Cause Category"]].value_counts()
```

```
Cause Category
MATERIAL/WELD/EQUIP FAILURE    1435
CORROSION                       592
INCORRECT OPERATION             378
ALL OTHER CAUSES                118
NATURAL FORCE DAMAGE            118
EXCAVATION DAMAGE                97
OTHER OUTSIDE FORCE DAMAGE       57
```


Provide Different Summary Stats

```
df_sample.describe()
```

	Net Loss (Barrels)	All Costs
count	2795.000000	2.795000e+03
mean	132.194050	8.340332e+05
std	1185.019252	1.657830e+07
min	0.000000	0.000000e+00
25%	0.000000	5.039500e+03
50%	0.000000	2.312900e+04
75%	2.000000	1.172325e+05
max	30565.000000	8.405261e+08

Count unique categories

```
df_sample.nunique()
```

```
Pipeline Location          2
Pipeline Type              4
Cause Category             7
Cause Subcategory         38
Net Loss (Barrels)       443
All Costs               2279
```

Check missing values

```
df_sample.isna().sum()
```

```
Pipeline Location       0
Pipeline Type          18
Cause Category          0
Cause Subcategory       0
Net Loss (Barrels)      0
All Costs               0
```

Show frequency of unique values

```
df_sample[["Cause Category"]].value_counts()
```

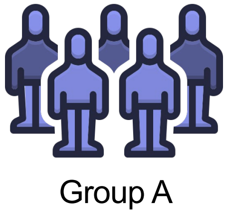
```
Cause Category
MATERIAL/WELD/EQUIP FAILURE    1435
CORROSION                       592
INCORRECT OPERATION             378
ALL OTHER CAUSES                118
NATURAL FORCE DAMAGE            118
EXCAVATION DAMAGE                97
OTHER OUTSIDE FORCE DAMAGE       57
```

Provide Different Summary Stats

```
df_sample.describe()
```

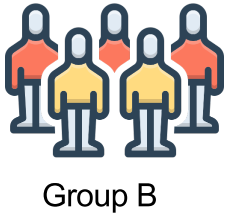
	Net Loss (Barrels)	All Costs
count	2795.000000	2.795000e+03
mean	132.194050	8.340332e+05
std	1185.019252	1.657830e+07
min	0.000000	0.000000e+00
25%	0.000000	5.039500e+03
50%	0.000000	2.312900e+04
75%	2.000000	1.172325e+05
max	30565.000000	8.405261e+08

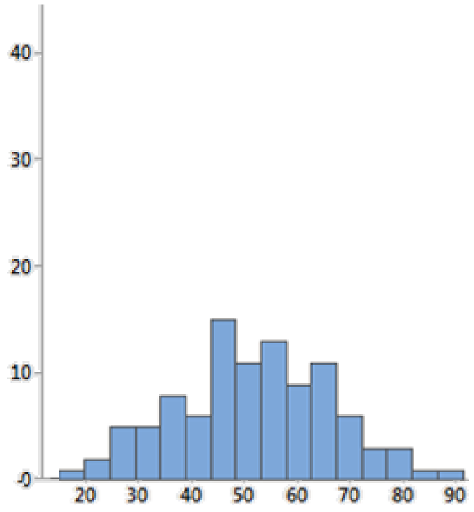
Summary Statistics provide a quick overview of your dataset

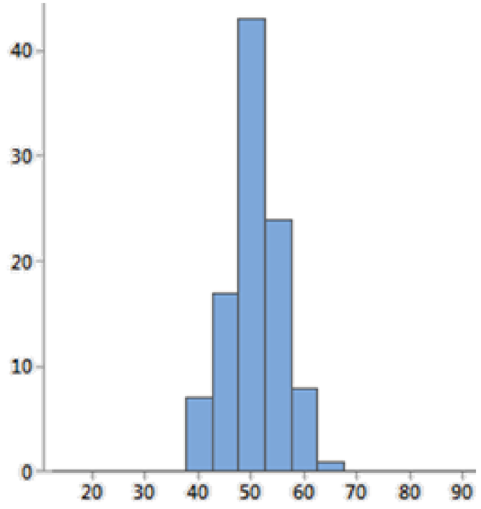


There are 2 classes of students going through same Math test

Math Test







Central Tendency shows the typical or central value of dataset

Dataset on house prices

House Number	House Price (AED)
101	1,200,000
102	1,100,000
103	1,250,000
104	1,100,000
105	1,200,000
106	1,300,000
107	1,150,000
108	1,300,000
109	1,400,000
110	3,500,000

- If you had to report the typical house price from this dataset, what value would you choose?

You can use a measure of central tendency – mean, median, or mode

Measures of central tendency include mean, median and mode

House Number	House Price (AED)
101	1,200,000
102	1,100,000
103	1,250,000
104	1,200,000
105	1,200,000
106	1,300,000
107	1,150,000
108	1,300,000
109	1,400,000
110	3,500,000

Mean

Sum all values
1,200,000+1,100,000+1,250,000+1,200,000+1,200,000+1,300,000+1,150,000+1,300,000+1,400,000+3,500,000=15,600,000

Divide by number of values
15,600,000/10 = 1,560,000

Measures of central tendency include mean, median and mode

House Number	House Price (AED)
101	1,200,000
102	1,100,000
103	1,250,000
104	1,200,000
105	1,200,000
106	1,300,000
107	1,150,000
108	1,300,000
109	1,400,000
110	3,500,000

Median

First, arrange in order

Measures of central tendency include mean, median and mode

House Number	House Price (AED)
102	1,100,000
107	1,150,000
101	1,200,000
104	1,200,000
105	1,200,000
103	1,250,000
106	1,300,000
108	1,300,000
109	1,400,000
110	3,500,000

Median

First, arrange in order

Then, find the middlemost value
(1,200,000+1,250,000)/2 = 1,225,000

Measures of central tendency include mean, median and mode

House Number	House Price (AED)
101	1,200,000
102	1,100,000
103	1,250,000
104	1,200,000
105	1,200,000
106	1,300,000
107	1,150,000
108	1,300,000
109	1,400,000
110	3,500,000

Mode

Find the value that occurs the most
Here, 1,200,000 appears thrice

Test your knowledge!

Which would help you identify the month with the most holidays?

A. Mean

B. Median

C. Mode

Test your knowledge!

Which would help you identify the month with the most holidays?

A. Mean

B. Median

C. Mode

Measures of dispersion show how spread out your data is

House Number	House Price (AED)
101	1,200,000
102	1,100,000
103	1,250,000
104	1,200,000
105	1,200,000
106	1,300,000
107	1,150,000
108	1,300,000
109	1,400,000
110	3,500,000

Range

Difference between largest and smallest values

$$Range(x) = \max(x) - \min(x)$$

Here, range would be

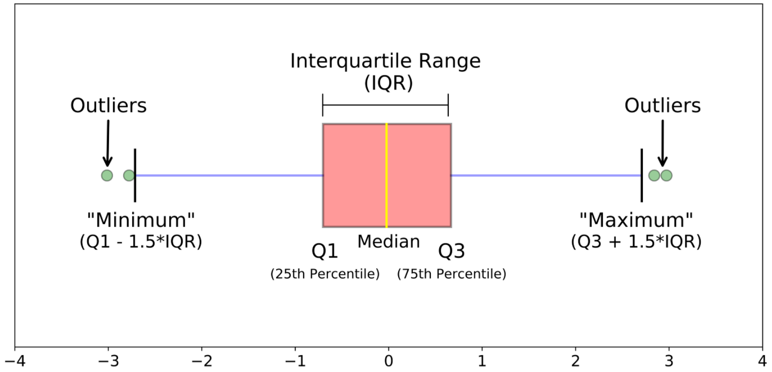
(3,500,000-1,100,000) = 2,400,000

Simply shows the full spread between smallest and largest values

Measures of dispersion show how spread out your data is

House Number	House Price (AED)
102	1,100,000
107	1,150,000
101	1,200,000
104	**1,200,000**
105	1,200,000
103	1,250,000
106	1,300,000
108	1,300,000
109	1,400,000
110	3,500,000

Inter-quartile Range



Difference between 25^{th} % and 75^{th} % of the values

$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}$$

Shows the spread of middle 50% of data while ignoring **outliers**

Measures of dispersion show how spread out your data is

House Number	House Price (AED)
102	1,100,000
107	1,150,000
101	1,200,000
104	**1,200,000**
105	1,200,000
103	1,250,000
106	1,300,000
108	1,300,000
109	1,400,000
110	3,500,000

Inter-quartile Range

Here, inter-quartile range would be

IQR=Q3−Q1=1,300,000−1,200,000=100,000

Shows the spread of middle 50% of data while ignoring **outliers**

Measures of dispersion show how spread out your data is

$$Standard\ Deviation = Square\ Root\ of\ \sum_{i=1}^{N}(X_i\ - Mean)^2$$

House Number	House Price (AED)
101	1320000
102	1230000
103	1350000
104	1480000
105	1210000
106	1210000
107	1490000
108	1370000
109	1180000
110	1330000
Std Dev	104790

House Number	House Price (AED)
101	1000000
102	1000000
103	1400000
104	500000
105	500000
106	900000
107	700000
108	1500000
109	700000
110	500000
Std Dev	343656

Mean is 1275000 for both

Shows how spread-out data points are around the average, here, 1275000

Missing values can bias results and reduce model accuracy

Quick Recap

Checking for missing values

Command

```
print(df.isnull().sum())
✓ 0.0s
```

Output

```
Date              0
Oil_Production    4
Gas_Production    4
Water_Cut         3
Field             0
dtype: int64
```

Handling Missing Values

Drop rows with missing values

Fill missing values with 0

Fill missing values with column mean

df.fillna (0, inplace=True)

df.dropna (inplace=True)

df [“Oil_Output”].fillna (df [“Oil_Output”].mean(), inplace=True)

Missing numerical values can be filled with numerical averages

Original

Command

Output

```
dff[['All Costs']]
```
✓ 0.0s

	All Costs
0	16169.0
1	16025.0
2	NaN
3	40000.0
4	17485.0

```
dff[['All Costs']]
    .fillna(dff['All Costs'].mean())
```
✓ 0.2s

Fill with mean or median

	All Costs
0	16169.00
1	16025.00
2	22419.75
3	40000.00
4	17485.00

Missing categorical values can be filled with mode or placeholder

Original → **Command** → **Output**

```
dff[['Pipeline Type']]
```
✓ 0.0s

	Pipeline Type
0	ABOVEGROUND
1	ABOVEGROUND
2	NaN
3	UNDERGROUND
4	UNDERGROUND

```
dff[['Pipeline Type']]
    .fillna(dff['Pipeline Type']
    .mode()[0])
```
✓ 0.0s

Fill with mode

	Pipeline Type
0	ABOVEGROUND
1	ABOVEGROUND
2	ABOVEGROUND
3	UNDERGROUND
4	UNDERGROUND

```
dff[['Pipeline Type']]
    .fillna('Unknown')
```
✓ 0.0s

Add placeholder

	Pipeline Type
0	ABOVEGROUND
1	ABOVEGROUND
2	Unknown
3	UNDERGROUND
4	UNDERGROUND

Handling Outliers is essential to prepare your dataset

🚨 **What is an Outlier?**

An **outlier** is a data point that deviates ***significantly*** from the rest of the dataset.

House Number	House Price (AED)
101	1,200,000
102	1,100,000
103	1,250,000
104	1,200,000
105	1,200,000
106	1,300,000
107	1,150,000
108	1,300,000
109	1,400,000
110	3,500,000

Remember this house that was pulling up your mean? That's your outlier!

Handling Outliers is essential to prepare your dataset

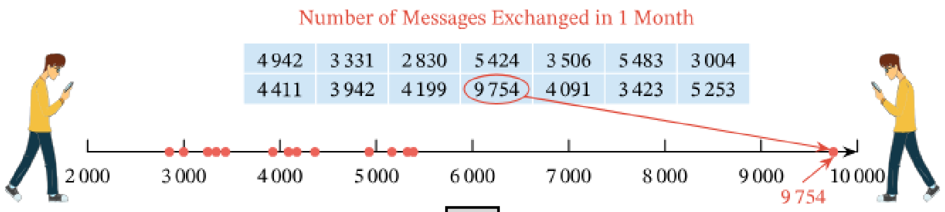
🚨 **What is an Outlier?**

An **outlier** is a data point that deviates ***significantly*** from the rest of the dataset.

House Number	House Price (AED)
101	1,200,000
102	1,100,000
103	1,250,000
104	1,200,000
105	1,200,000
106	1,300,000
107	1,150,000
108	1,300,000
109	1,400,000
110	3,500,000

How do we decide if the deviation is significant?

Outliers are those values which lie outside the inter-quartile range



	Count	Mean	Std	Min	25%	50%	75%	Max
Values	14	4399	1737	2830	3438	4016	4809	9754

Metric	Calculation	Value
Lower Bound	Q1 - (1.5 × IQR)	1381
Upper Bound	Q3 + (1.5 × IQR)	6866

Outliers	Values outside Lower & Upper Bound	9754

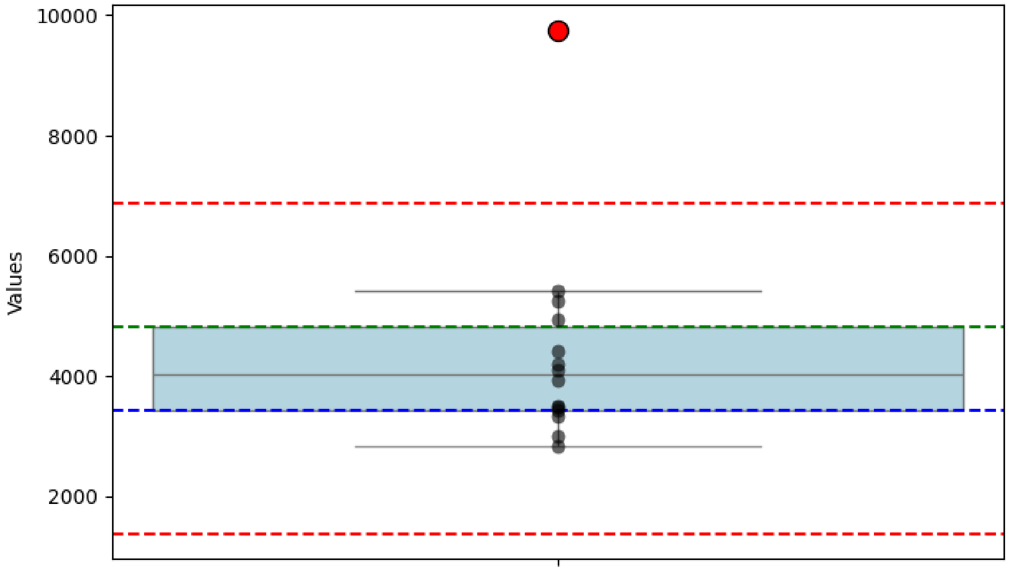
Outliers can be rectified easily with the help of Python

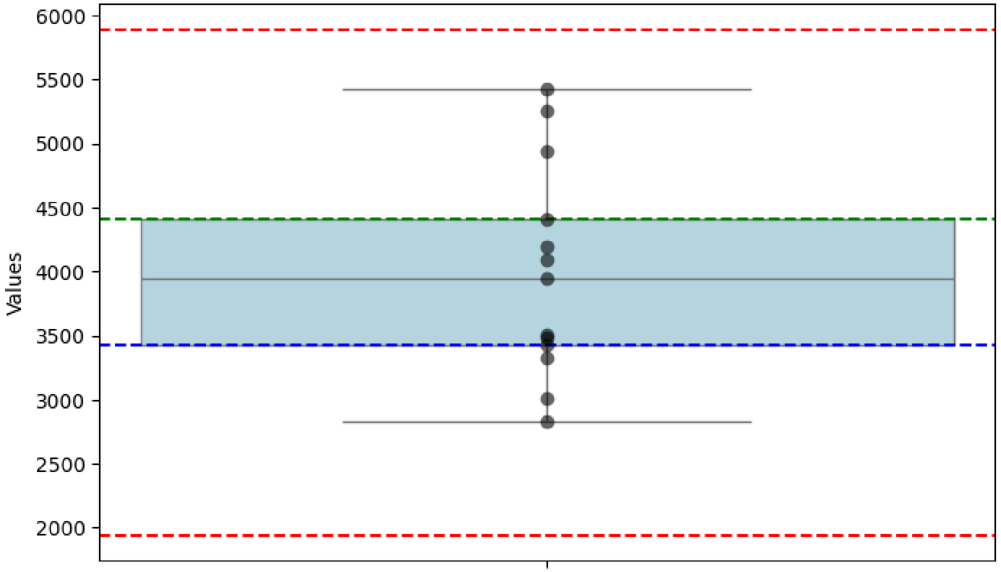
Fix	Command
Remove extreme values	`df = df[df['column_name'] < threshold]` (e.g., using IQR method or Z-score)
Replace with mean or median	`df['column_name'].replace(outlier_value, df['column_name'].median())`

Outliers can be rectified easily with the help of Python

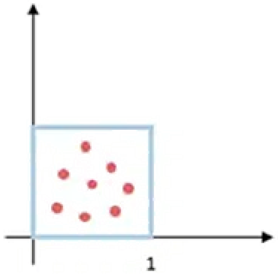
Before Fix

After Fix





Data can be rescaled for better outlier handling



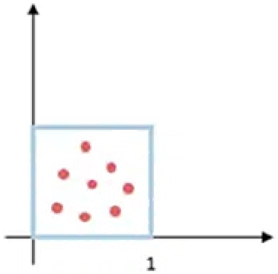
House Number	House Price (AED)
101	1,200,000
102	1,100,000
103	1,250,000
104	1,200,000
105	1,200,000
106	1,300,000
107	1,150,000
108	1,300,000
109	1,400,000
110	3,500,000

Min-Max Scaling

Scales values between 0 and 1

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Data can be rescaled for better outlier handling



House Number	House Price (Normalised)
101	0.0417
102	0
103	0.0625
104	0.0417
105	0.0417
106	0.0833
107	0.0208
108	0.0833
109	0.125
110	1

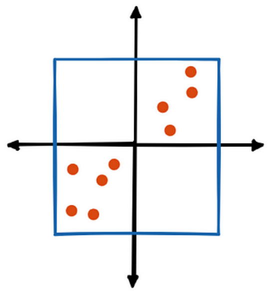
Min-Max Scaling

Scales values between 0 and 1

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Best for data within a range, free from outliers,
and preserving relative distances

Data can be normalised to scale for better outlier handling



House Number	House Price (AED)
101	1,200,000
102	1,100,000
103	1,250,000
104	1,200,000
105	1,200,000
106	1,300,000
107	1,150,000
108	1,300,000
109	1,400,000
110	3,500,000

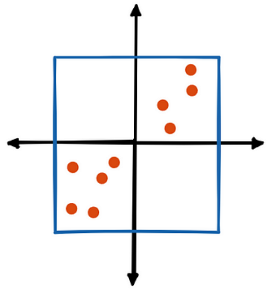
Standard Scaling

Transform data to have zero mean and unit variance

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

Data can be normalised to scale for better outlier handling



House Number	House Price (Standardised)
101	-0.380
102	-0.526
103	-0.307
104	-0.380
105	-0.380
106	-0.234
107	-0.453
108	-0.234
109	-0.088
110	1

Standard Scaling

Transform data to have zero mean and unit variance

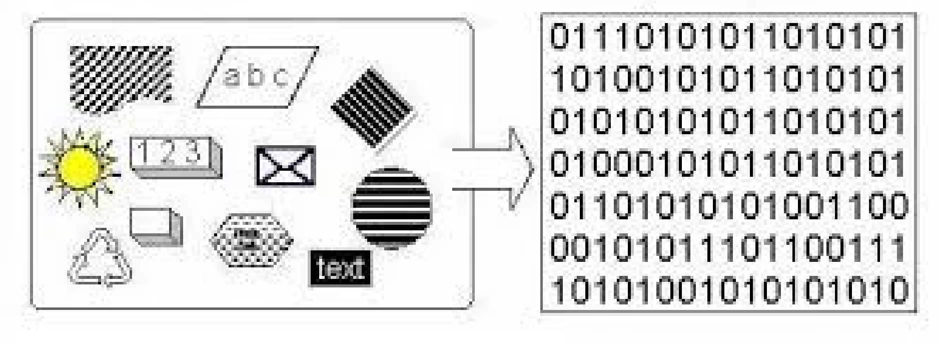
$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

This method is used often,
especially for data that assume normal distribution

Feature Engineering can make input data suitable for ML

Input Data to a Machine Learning Model



What Computer understands ??

Ordinal Data can be treated with Label Encoding

Label Encoding (Integer Encoding)
Replaces each unique category with an integer

Index	Pipeline Type
0	ABOVEGROUND
1	UNDERGROUND
2	ABOVEGROUND
3	ABOVEGROUND
4	TANK

Index	Pipeline Type Encoded
0	0
1	1
2	0
3	0
4	2

Nominal categories are best handled with One-Hot Encoding

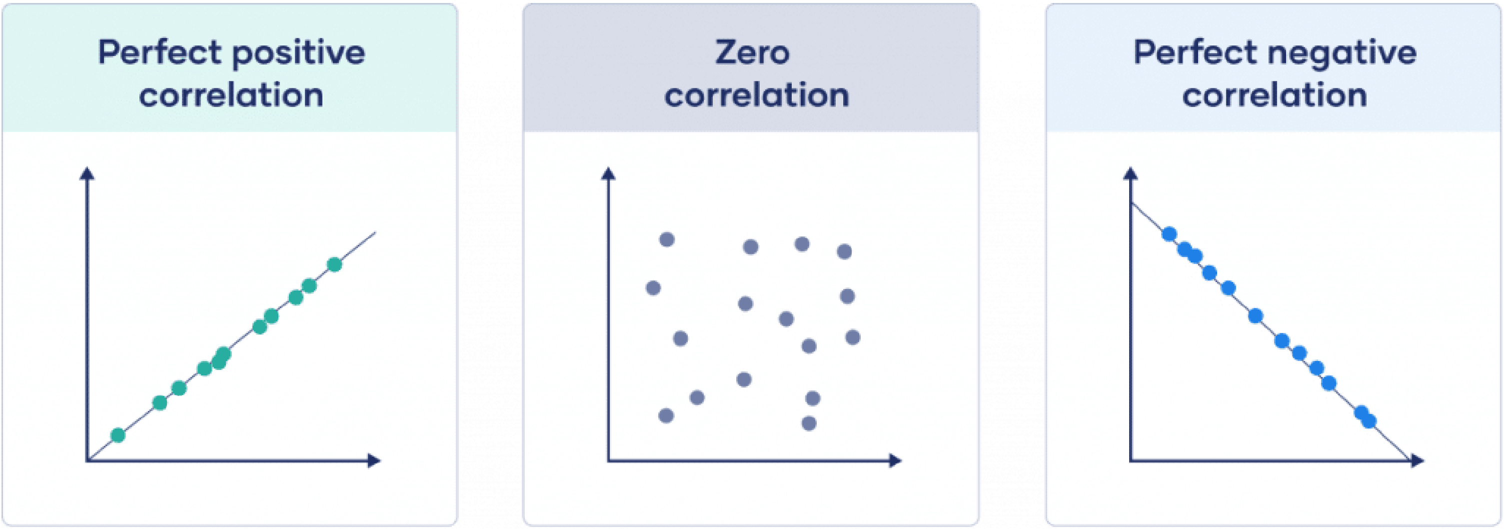
One-Hot Encoding (OHE)
Creates a **binary column** for each category

Index	Pipeline Type
0	ABOVEGROUND
1	UNDERGROUND
2	ABOVEGROUND
3	ABOVEGROUND
4	UNDERGROUND

Index	Pipeline Type Aboveground	Pipeline Type Underground
0	1	0
1	0	1
2	1	0
3	1	0
4	0	1

Correlation issues can destabilise and slow down your model

Correlation is simply the relationship between two variables



Redundant information

Slower Learning

Misleading Results

Some features you chose may be related to each other

Pipeline Location	Pipeline Type	Cause Category	Cause Subcategory	Net Loss (Barrels)	All Costs
ONSHORE	UNDERGROUND	CORROSION	INTERNAL	0.00	5065
ONSHORE	UNDERGROUND	MATERIAL/WELD/EQUIP FAILURE	CONSTRUCTION, INSTALLATION OR FABRICATION-RELATED	0.00	216121
ONSHORE	TANK	ALL OTHER CAUSES	MISCELLANEOUS	0.00	16200
ONSHORE	ABOVEGROUND	MATERIAL/WELD/EQUIP FAILURE	MALFUNCTION OF CONTROL/RELIEF EQUIPMENT	0.00	32477
ONSHORE	UNDERGROUND	INCORRECT OPERATION	DAMAGE BY OPERATOR OR OPERATOR'S CONTRACTOR	25.00	84600
ONSHORE	UNDERGROUND	EXCAVATION DAMAGE	THIRD PARTY EXCAVATION DAMAGE	0.00	709351
ONSHORE	ABOVEGROUND	INCORRECT OPERATION	INCORRECT VALVE POSITION	0.00	16169
ONSHORE	ABOVEGROUND	CORROSION	EXTERNAL	0.00	16025
OFFSHORE	NaN	CORROSION	EXTERNAL	0.10	1450000
ONSHORE	UNDERGROUND	MATERIAL/WELD/EQUIP FAILURE	CONSTRUCTION, INSTALLATION OR FABRICATION-RELATED	0.00	40000
ONSHORE	UNDERGROUND	CORROSION	INTERNAL	0.00	17485
ONSHORE	ABOVEGROUND	CORROSION	INTERNAL	0.36	5040
ONSHORE	ABOVEGROUND	INCORRECT OPERATION	OVERFILL/OVERFLOW OF TANK/VESSEL/SUMP	0.00	11607
ONSHORE	ABOVEGROUND	MATERIAL/WELD/EQUIP FAILURE	MALFUNCTION OF CONTROL/RELIEF EQUIPMENT	0.35	6350
ONSHORE	ABOVEGROUND	MATERIAL/WELD/EQUIP FAILURE	PUMP OR PUMP-RELATED EQUIPMENT	0.00	3770
ONSHORE	ABOVEGROUND	OTHER OUTSIDE FORCE DAMAGE	VEHICLE NOT ENGAGED IN EXCAVATION	0.10	150001
ONSHORE	UNDERGROUND	EXCAVATION DAMAGE	OPERATOR/CONTRACTOR EXCAVATION DAMAGE	0.00	40750
ONSHORE	ABOVEGROUND	OTHER OUTSIDE FORCE DAMAGE	VEHICLE NOT ENGAGED IN EXCAVATION	976.00	776753
ONSHORE	ABOVEGROUND	MATERIAL/WELD/EQUIP FAILURE	NON-THREADED CONNECTION FAILURE	3.00	1270
ONSHORE	ABOVEGROUND	INCORRECT OPERATION	INCORRECT INSTALLATION	0.35	2034

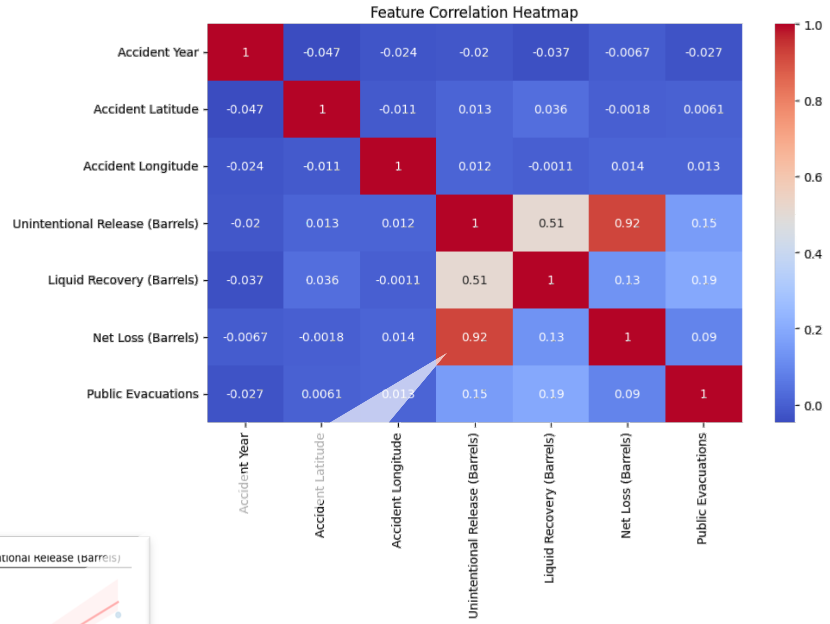
Remember this dataset we saw?

Handling correlation is necessary for your model to be reliable

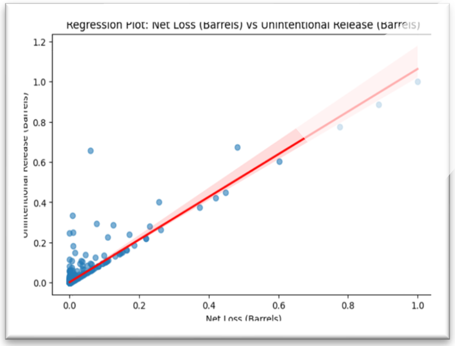
Let’s say you chose two features to incorporate in you model:
Unintentional Release and Net Loss

Handling correlation is necessary for your model to be reliable

These two features 'Unintentional Release' and 'Net Loss' show a very high correlation

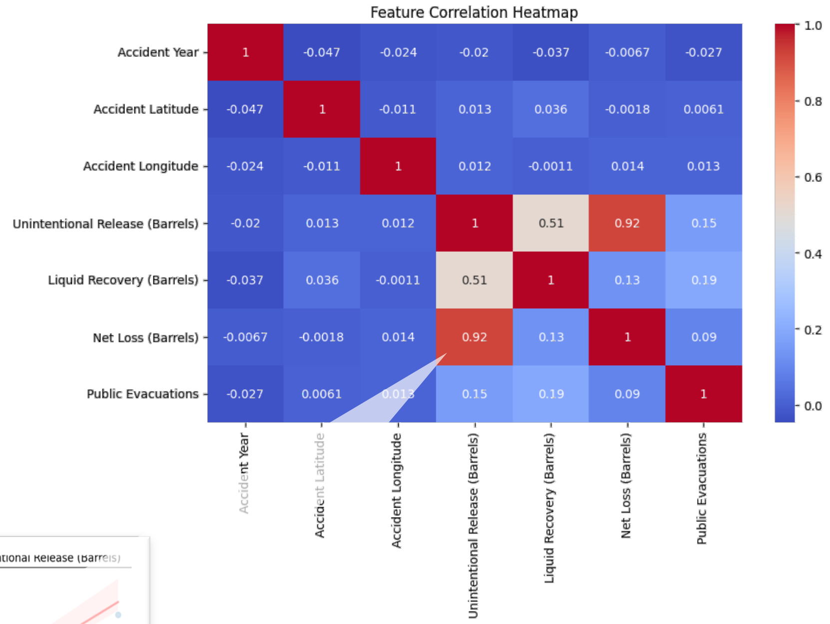


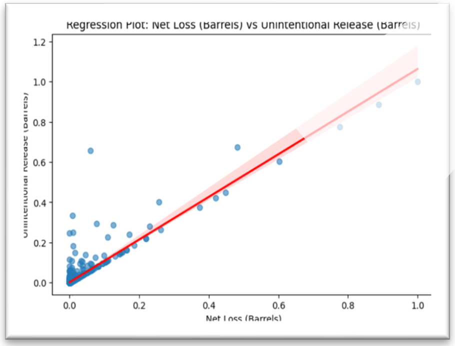
	Accident Year	Accident Latitude	Accident Longitude	Unintentional Release (Barrels)	Liquid Recovery (Barrels)	Net Loss (Barrels)	Public Evacuations
Accident Year	1	-0.047	-0.024	-0.02	-0.037	-0.0067	-0.027
Accident Latitude	-0.047	1	-0.011	0.013	0.036	-0.0018	0.0061
Accident Longitude	-0.024	-0.011	1	0.012	-0.0011	0.014	0.013
Unintentional Release (Barrels)	-0.02	0.013	0.012	1	0.51	0.92	0.15
Liquid Recovery (Barrels)	-0.037	0.036	-0.0011	0.51	1	0.13	0.19
Net Loss (Barrels)	-0.0067	-0.0018	0.014	0.92	0.13	1	0.09
Public Evacuations	-0.027	0.0061	0.013	0.15	0.19	0.09	1



Handling correlation is necessary for your model to be reliable

Try to keep only one of the correlated features!





Redundant information

Slower Learning

Misleading Results

Data Transformation and Cleaning

In this session, we covered:

- Identifying and understanding data types
- Using summary statistics to describe a dataset
- Cleaning data by handling missing values and outliers
- Transforming data using normalisation and standardisation
- Using feature engineering to improve model training