Data Analytics

Mrinal Das

Learning Objective

Data Science and Analytics for Manufacturing

- Data Analysis
- Decision making
- Data processing
- Visualization of Analysis
- Statistical Analytics
- Hypothesis testing
- Regression analysis
- Design of Experiments
- Association Rule Mining
- Clustering

Basic Concepts, Examples, Hands-on Demonstrations

What is Data Analytics

Data analytics involves

the research, discovery, and interpretation of patterns

within data.

Moreover, we applied data analytics to solve problems.

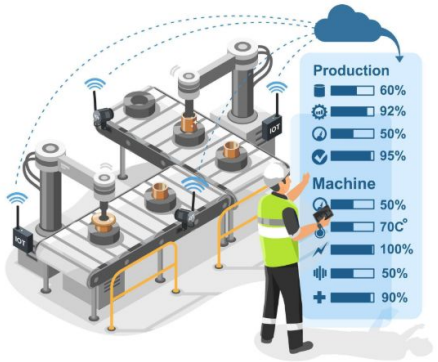
Goal of Data Analytics

Scientific ways to store, process, extract data

Data holds history, but hard to understand

Analysing data can reveal

- Cause of failure
- Possible future
- Ways to make future better



Important Components

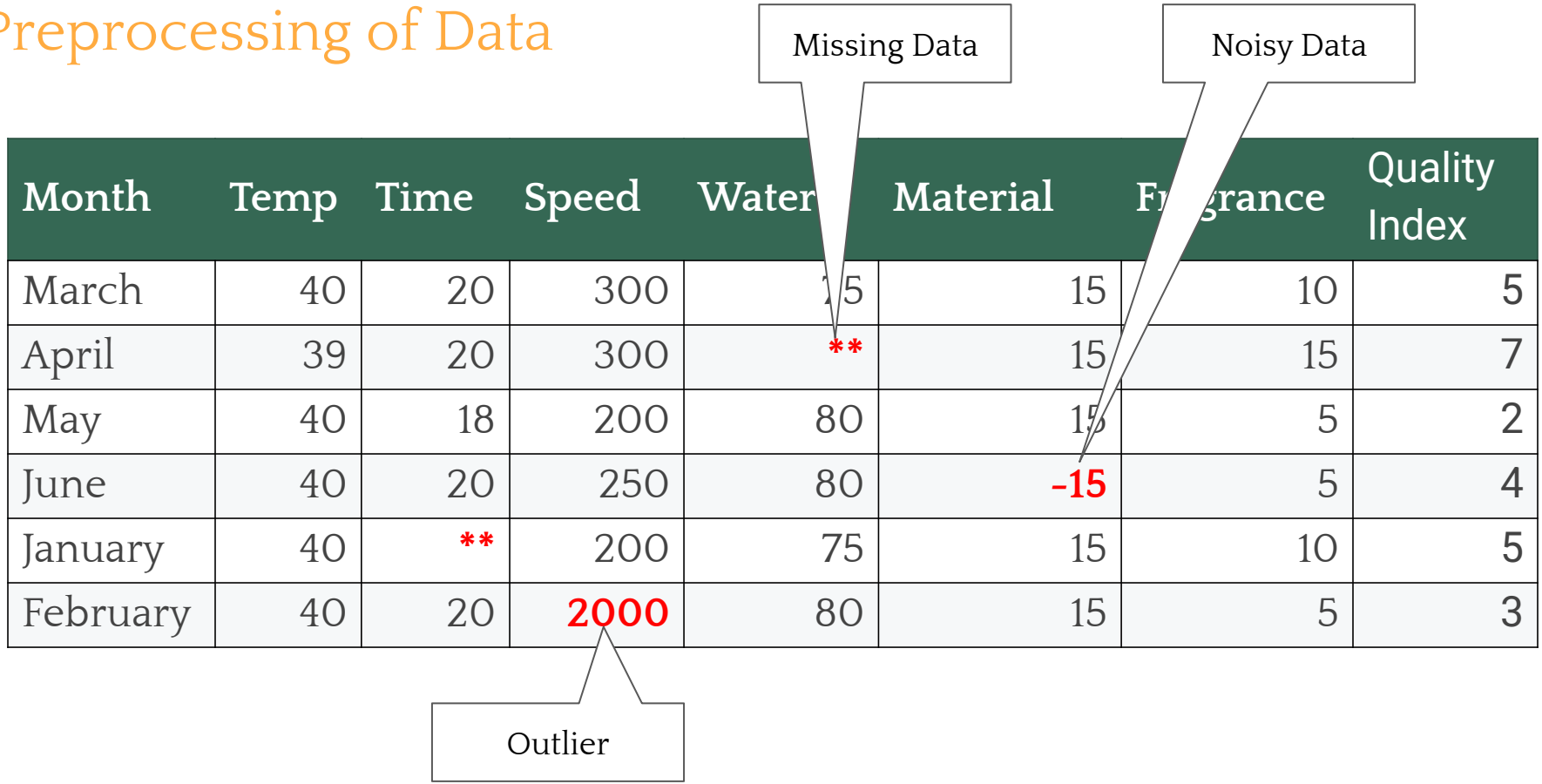
- ★ Read data
- ★ Clean, slice, massage, visualize
- ★ Compute statistical properties
- ★ Represent findings to be understood by all
- ★ Come up with remedial measures

- ★ Computer Program
- ★ Mathematics
- ★ Presentation Skill

Major Types of Analytics

Descriptive	Diagnostic
What happened	Why happened
Predictive	Prescriptive
What will happen	What to do

Preprocessing of Data



Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	**	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	-15	5	4
January	40	**	200	75	15	10	5
February	40	20	2000	80	15	5	3

Preprocessing to Make Data Ready

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

Visualization for Story Telling

Principle 1: Know your audience

Managers - Dashboard

Engineers - Detailed plots

Customers - Simple charts

Principle 2: Simplicity

Simple, clear, focused

Too many colors, too many charts, too many labels

Principle 3: Highlight

What is the point

Principle 4: Accuracy

Precise scales, numbers

3 Key Aspects of Data

Central Tendency

What is a typical value for each variable?

Dispersion

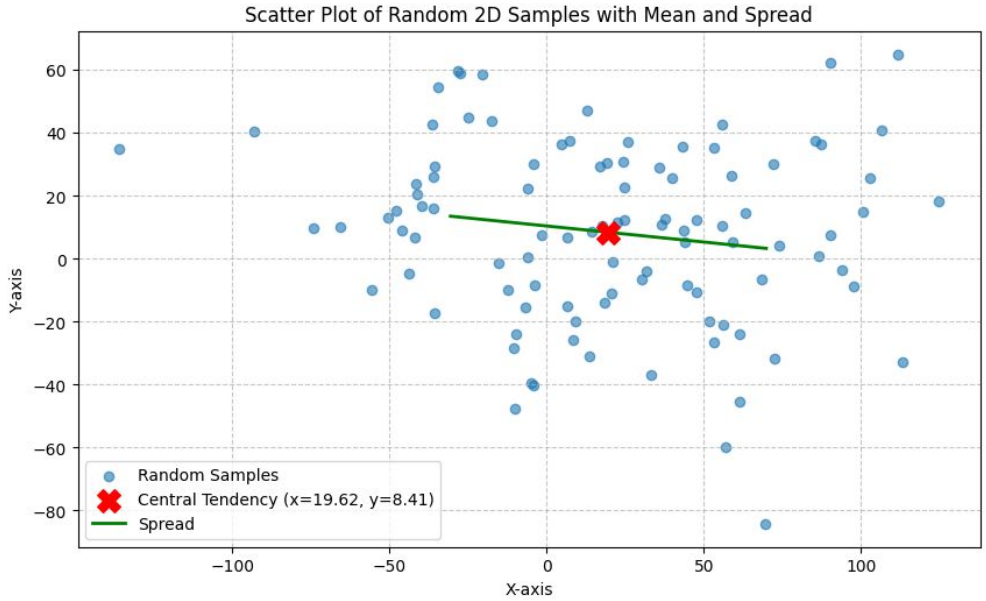
How far apart are the individual observations from a central value for a given variable?

Association

When more than one variable are studied together, how does each variable relate to the remaining variables?

Central Tendency and Dispersion

How far apart are the individual observations from a central value for a given variable?



Association

Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	70	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	15	5	4
January	40	18	200	75	15	10	5
February	40	20	200	80	15	5	3

Mean(water)=76.6, Mean(QI)=4.3.

Var(water)=16.6, Var(QI)=3.1

Cov(water,QI) = -5.56, Corr = -0.93

Mean(Fragrance)=8.3, Mean(QI)=4.3.

Var(Fragrance)=16.6, Var(QI)=3.1

Cov(Fragrance,QI) = 5.56, Corr = 0.93

Predictive Analytics using Probability

$P(QI > 4.5 \mid Water < 78) = ?$

We can not find this.

$P(QI > 4.5 \mid Water < 78) = P(QI>4.5, Water<78)/P(Water<78)$

Water	Quality Index
75	5
70	7
80	2
80	4
75	5
80	3

$P(QI>4.5,Water<78) = 3/6$

$P(Water<78)=3/6$

$P(QI > 4.5 \mid Water < 78) = 1$

Data Driven Decision Making

One can take a decision by checking the specifications and using expertise

However, it is not possible in every case

If the machine is a new product, or the task is new

Often encountered in technology replacements such as automation

Data driven approach is safe and scientific

Data to Decision

```
X = np.array([70, 63, 72, 68]) # success rate of machine-1
Y = np.array([65, 67, 63, 70]) # success rate of machine-2
```

Can not conclude that machine-1 is better

```
X = np.array([70, 72, 72, 68]) # success rate of machine-1
```

Machine-1 is better

```
X = np.array([70, 72, 72, 68, 68, 68])
```

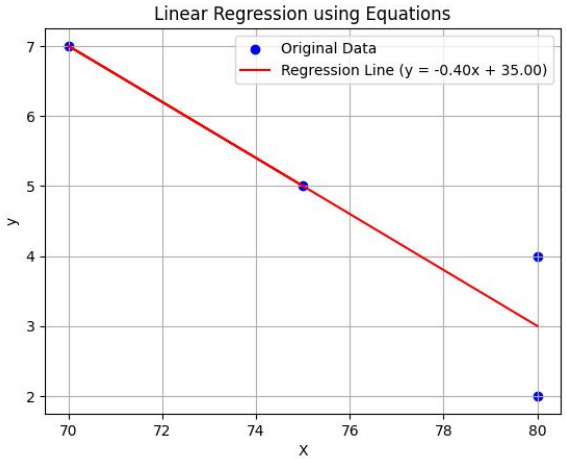
Machine-1 is better

Predictive Analytics using Regression

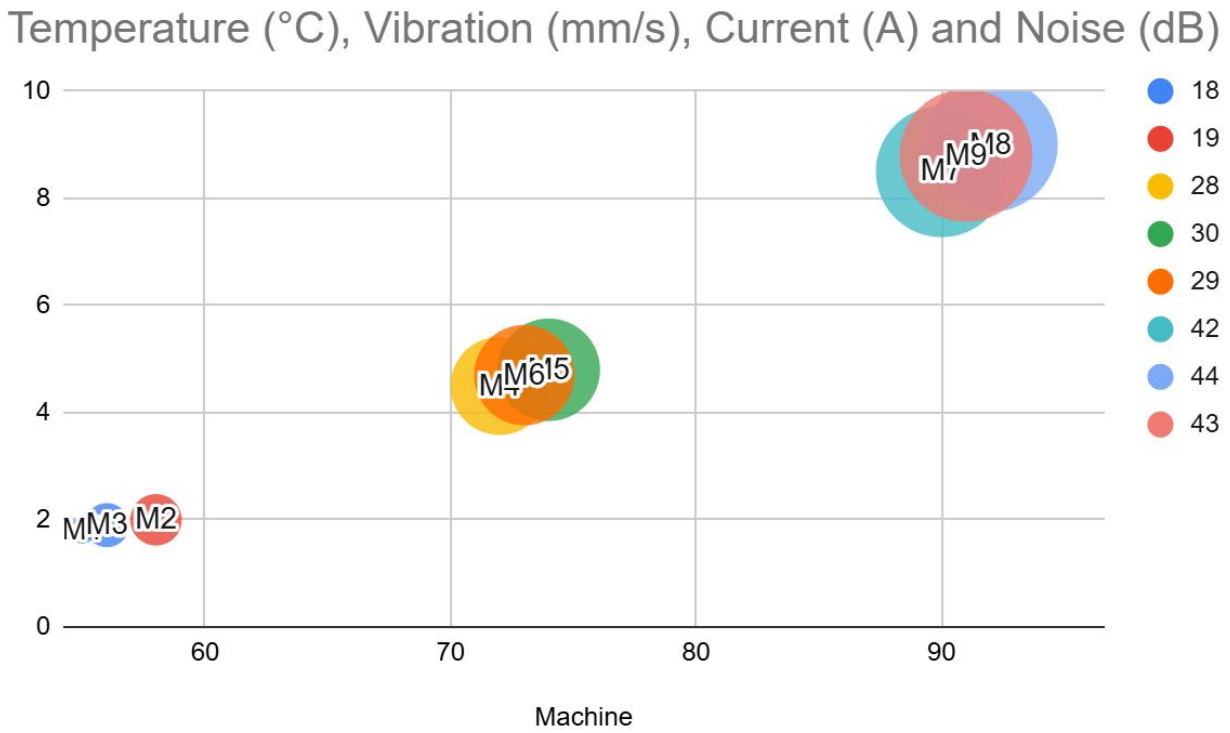
Water	Quality Index
75	5
70	7
80	2
80	4
75	5
80	3

QI = a * Water + b

Solution we get: a = -0.4, b=35



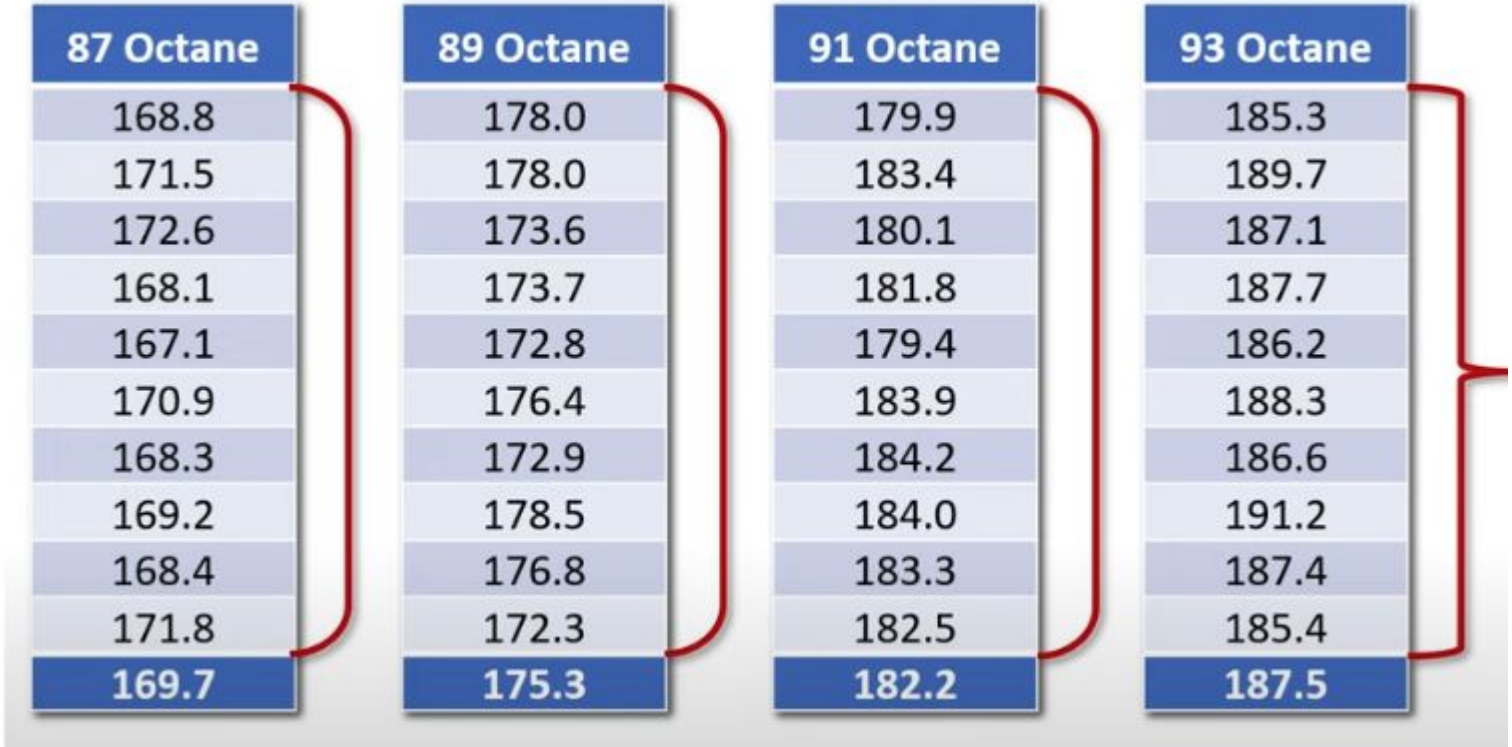
Descriptive Analytics using Clustering



Design of Experiments to Improve a Process

Suppliers	Inputs	Process	Outputs	Customers
***Providers** of Resources*	*Resources Required for Process*	*How we Transform inputs into outputs*	*Deliverables of Process*	*Receivers of Process Outputs*
Ingredient Vendors Grocery Store Home Goods Store (Equipment)	**Cake Ingredients** **Mixing Equipment** **Mixing Bowls** **Cake Pan** **The Oven** **Baking Temperature** **Baking Time** **Measuring Cup** **The Bakers**	1. Preheat Oven 2. **Mix Dry Ingredients** 3. **Mix Wet Ingredients** 4. Combine Wet and Dry 5. Pour into Pan 6. **Bake the cake** 7. Remove and cool 8. Decorate 9. Package and Ship	Cake **flavor** Cake **texture** Cake **color** Cake **size** Cake **shape** **Yield,** **Lead Time,** **Product Cost**	Bakery Customers Party Guests

ANOVA to Select Best Setting



87 Octane	89 Octane	91 Octane	93 Octane
168.8	178.0	179.9	185.3
171.5	178.0	183.4	189.7
172.6	173.6	180.1	187.1
168.1	173.7	181.8	187.7
167.1	172.8	179.4	186.2
170.9	176.4	183.9	188.3
168.3	172.9	184.2	186.6
169.2	178.5	184.0	191.2
168.4	176.8	183.3	187.4
171.8	172.3	182.5	185.4
169.7	175.3	182.2	187.5

Causality Analysis

Discover strong associations

$$HighTemp + HighVibration + LowLubrication \rightarrow BearingFailure$$

Suppose a new machine is installed and nobody has much experience with it.

Using data analytics one can find insights

Cause to Effect relationships

Helps in planning maintenance (reduce downtime, maintenance cost)

Case Study

Indian Chemicals has set up a data analytics team

The task is given to them :

"Use data to reduce defects, predict machine failures, and improve production?"

How will you do some similar project ?

Basic Information Collection

Indian Chemicals has 3 plants: plant A, plant B, plant C

Every day, the company collects thousands of records from:

- machines,
- sensors,
- quality inspection,
- production lines,
- maintenance systems,
- workers,
- customer orders.

Start with Raw Data Collection

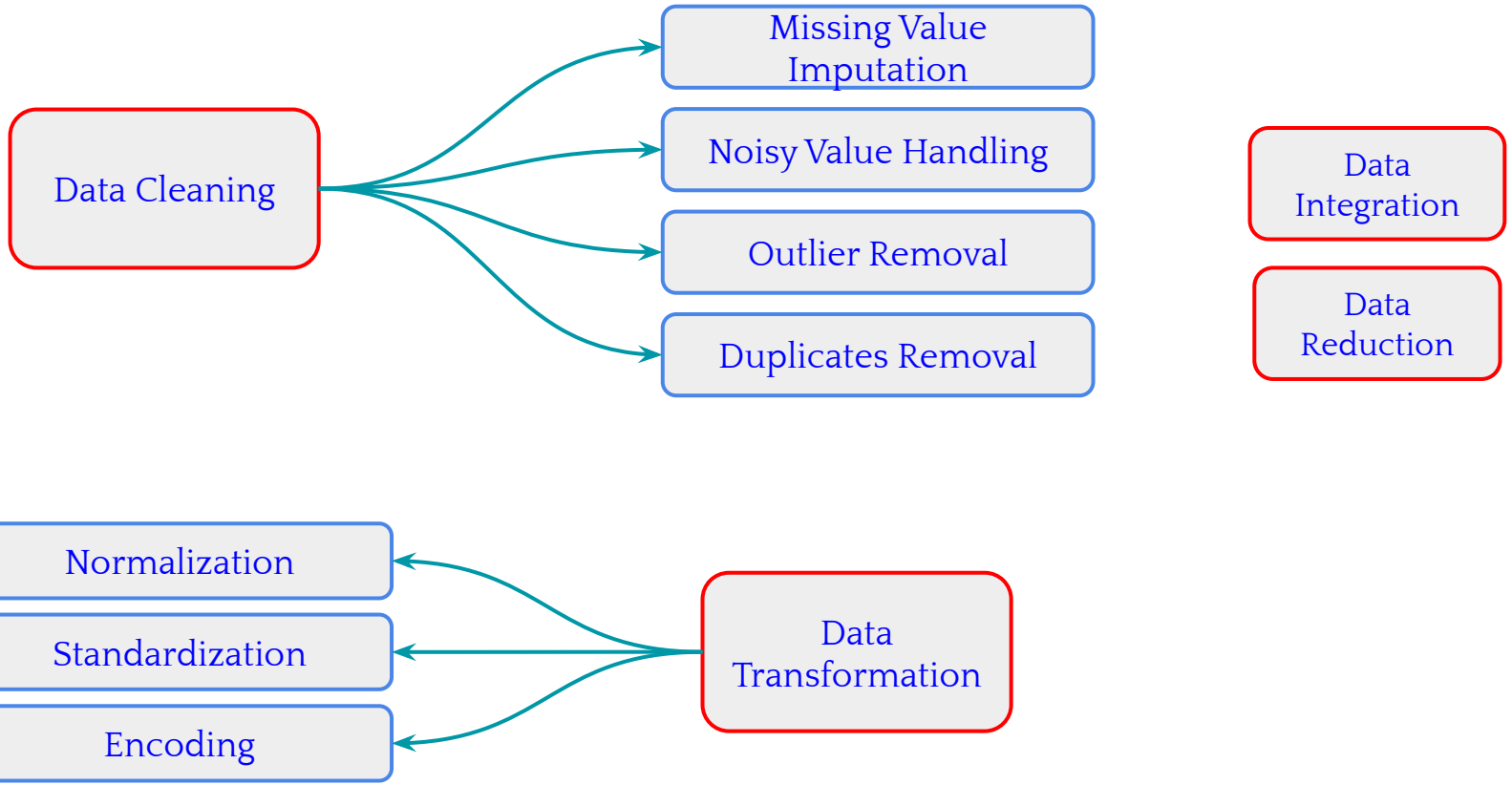
Product	Plant	Temperature	Vibration	Pressure	Speed	Weight	Defect
C001	A	72		101	80	1420	No
C002	A	81	5.8	104	85	1435	Yes
C003	B	69	2.9	0.99	78	1410	No
C004	C	84	6.2	106	87	a1450	Yes

Preprocess the Data

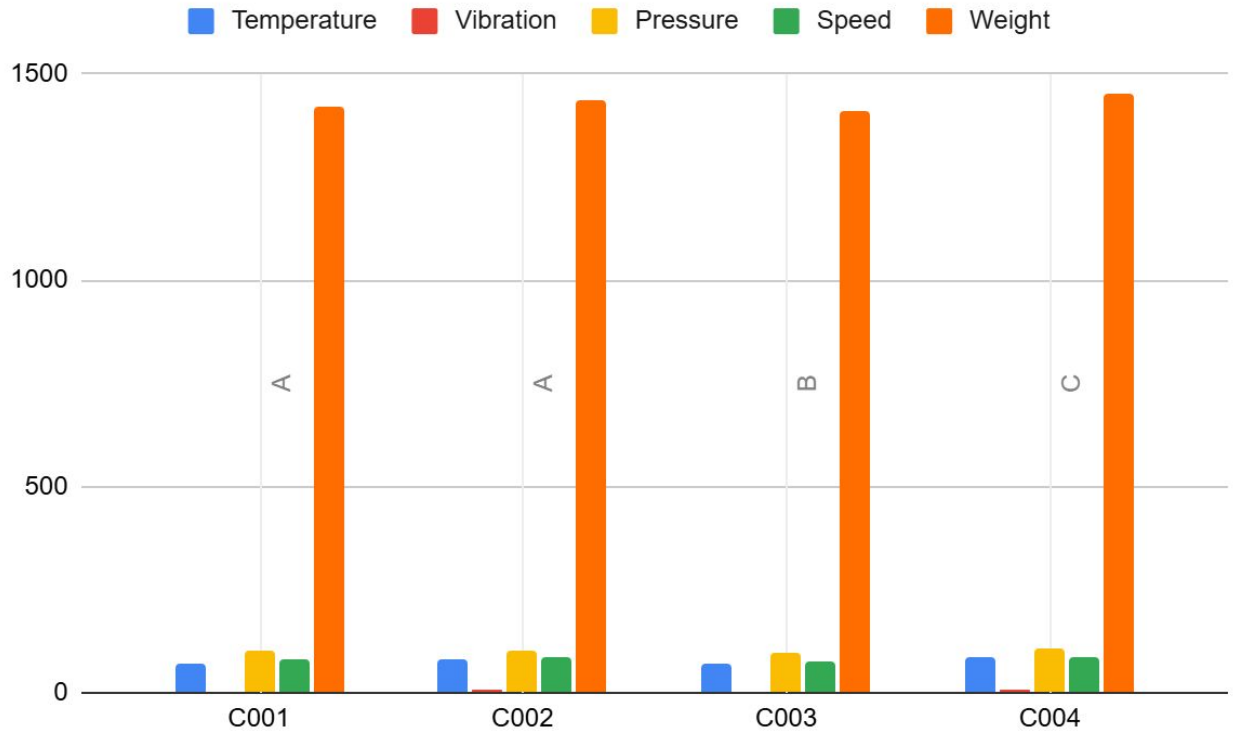
Product	Plant	Temperature	Vibration	Pressure	Speed	Weight	Defect
C001	A	72	**	101	80	1420	No
C002	A	81	5.8	104	85	1435	Yes
C003	B	69	2.9	0.99	78	1410	No
C004	C	84	6.2	106	87	a1450	Yes

missing values,
duplicate records,
inconsistent units,
outliers,
categorical variables,
noisy sensor readings.

Preprocessing

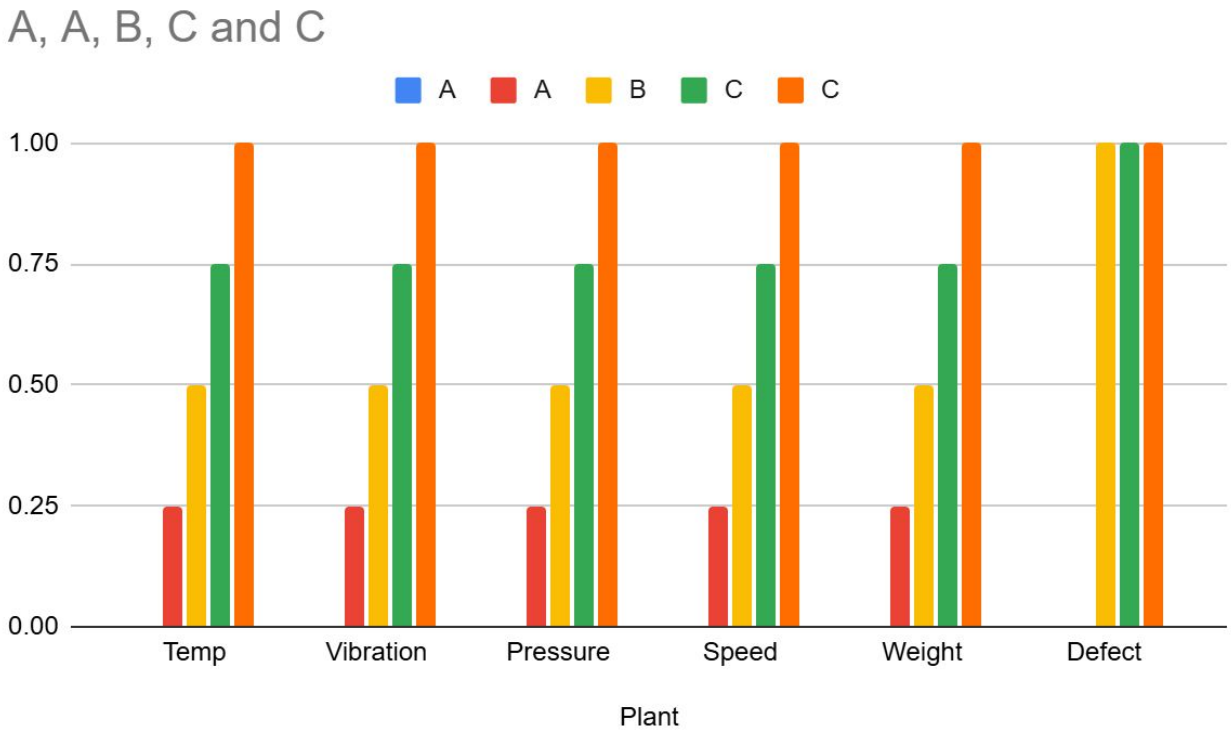


Preprocessing

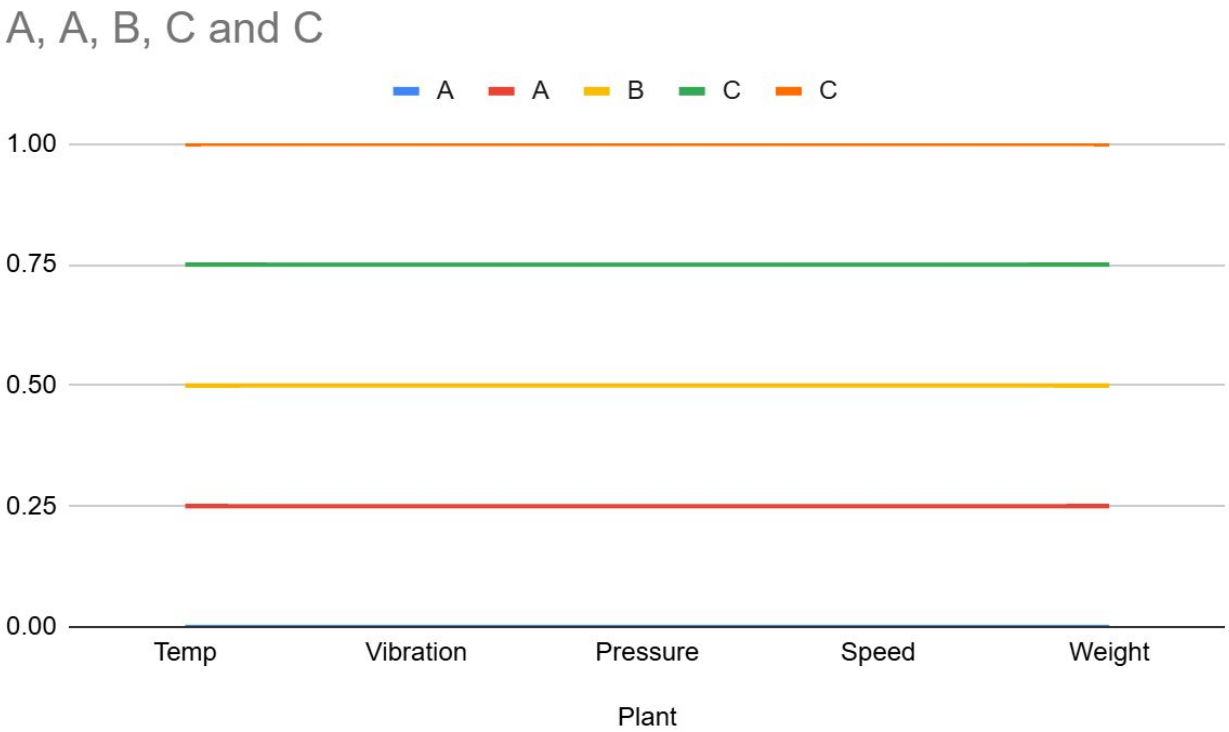


Cleaned Data

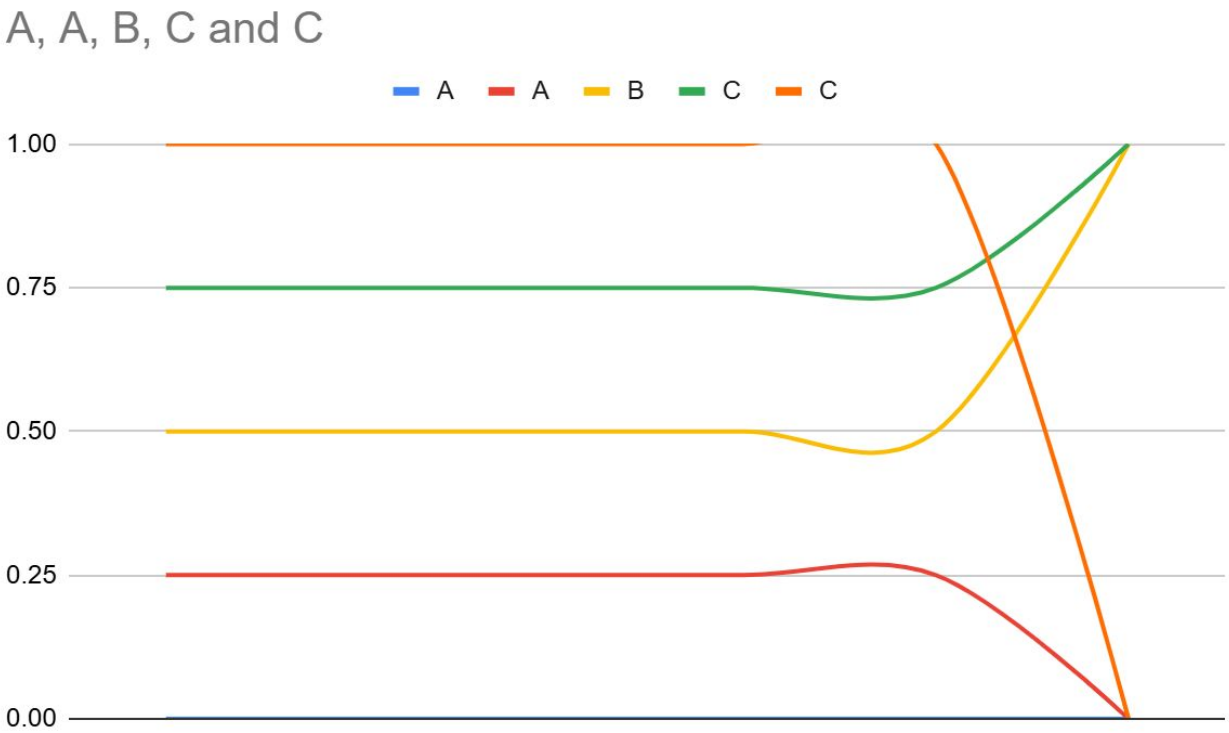
Plant	Temp	Vibration	Pressure	Speed	Weight	Defect
A	0	0	0	0	0	0
A	0.25	0.25	0.25	0.25	0.25	0
B	0.5	0.5	0.5	0.5	0.5	1
C	0.75	0.75	0.75	0.75	0.75	1
C	1	1	1	1	1	1


A, A, B, C and C
A
A
B
C
C
1.00
0.75
0.50
0.25
0.00
Temp
Vibration
Pressure
Speed
Weight
Defect
Plant

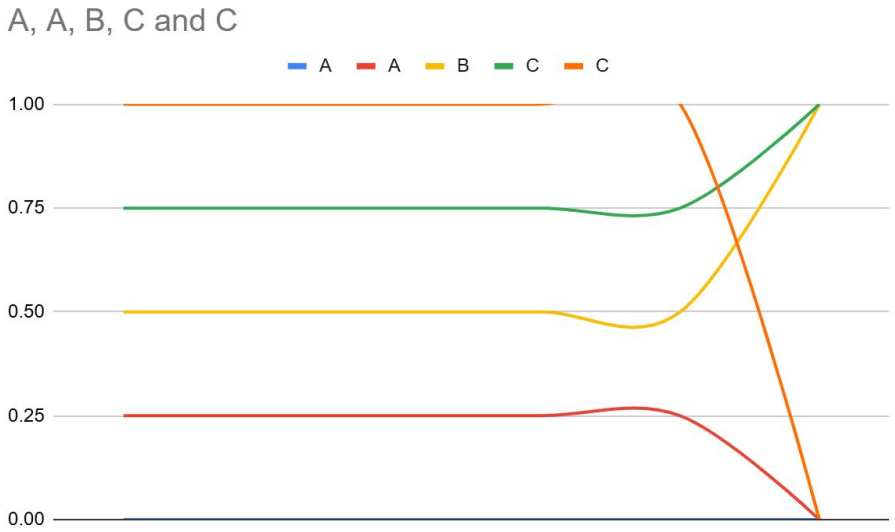
Plant wise View



Alternative Description of Data

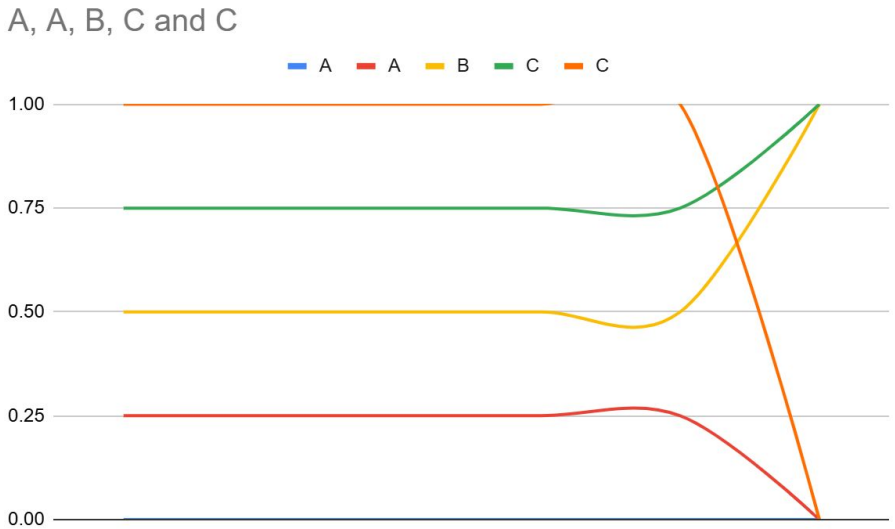


Alternative Description of Data



Why is defect rate high for plant B ?

Alternative Description of Data



Is it because of less sample and random effect ?

Significance Test

Suppose:

- Plant A defect rate = 4%
- Plant B defect rate = 7%

Is this difference statistically significant?

So we need to observe for longer time, collect more data. Then analyse.

A Different Angle

To investigate the issue with plant B, we can check if their parameters are very different from other plants.

We can do analysis of variance or ANOVA test.

We can also check association through correlation.

Diagnostic Analytics via Correlation

Feature	Correlation with Defect
Temperature	0.866
Vibration	0.866
Pressure	0.866
Speed	0.866
Weight	0.866

From this description we can not conclude anything.

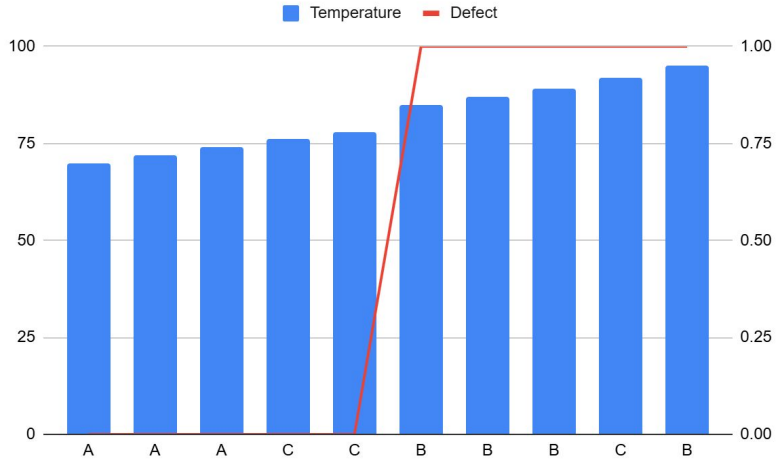
We may need more data.

Collect more Data

Plants	Temperature	Vibration	Pressure	Speed	Weight	Defect
A	70	7	105	82	1450	0
A	72	3	98	91	1390	0
A	74	8	112	68	1470	0
C	76	4	101	76	1410	0
C	78	6	109	95	1430	0
B	85	5	100	73	1460	1
B	87	9	115	88	1380	1
B	89	2	97	79	1490	1
C	92	7	106	67	1420	1
B	95	4	111	93	1400	1

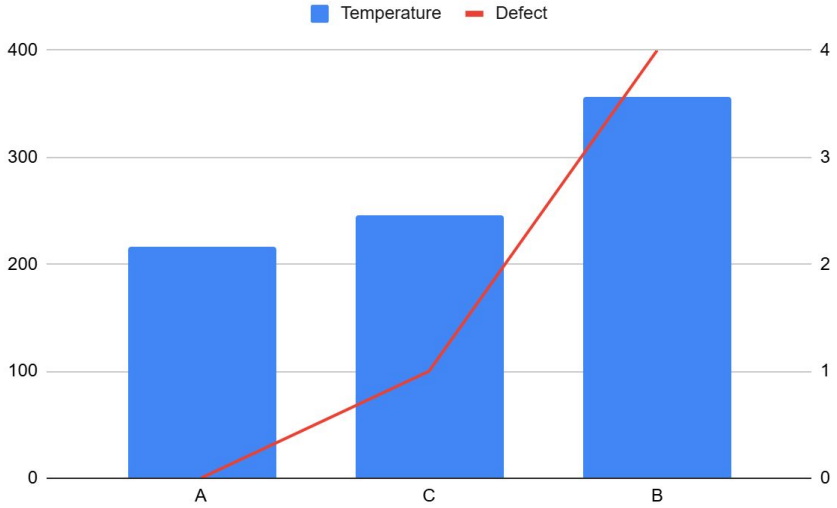
Repeat Diagnostic Analysis

Feature	Correlation with Defect
Temperature	≈ +0.98
Vibration	≈ −0.05
Pressure	≈ +0.13
Speed	≈ −0.02
Weight	≈ −0.03



Temperature has high correlation with Defects

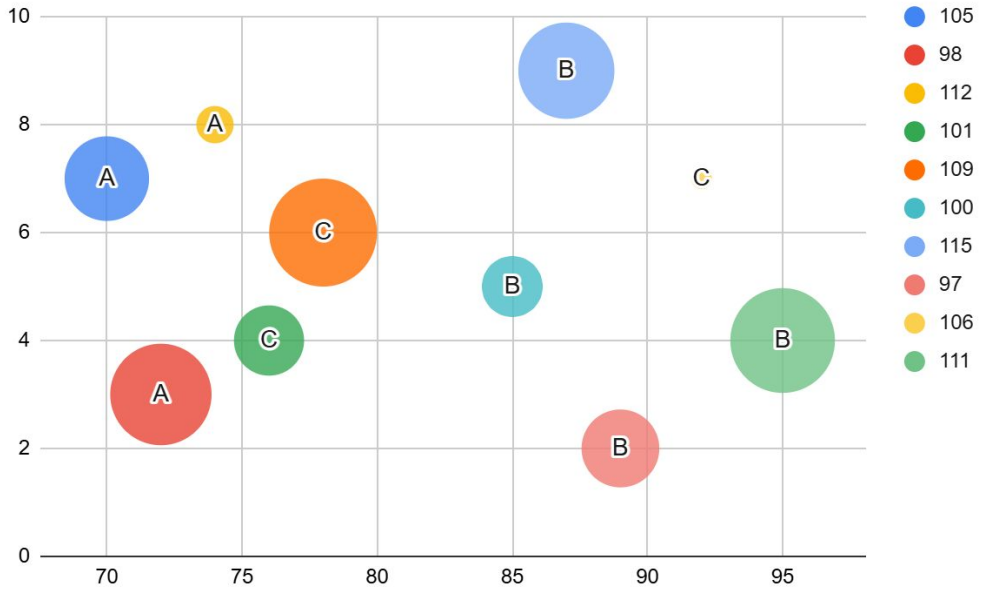
Customizing the Plots



It is more clear that plant B has more temperature and also more defects.

We can also apply Association Rule mining in such investigations.

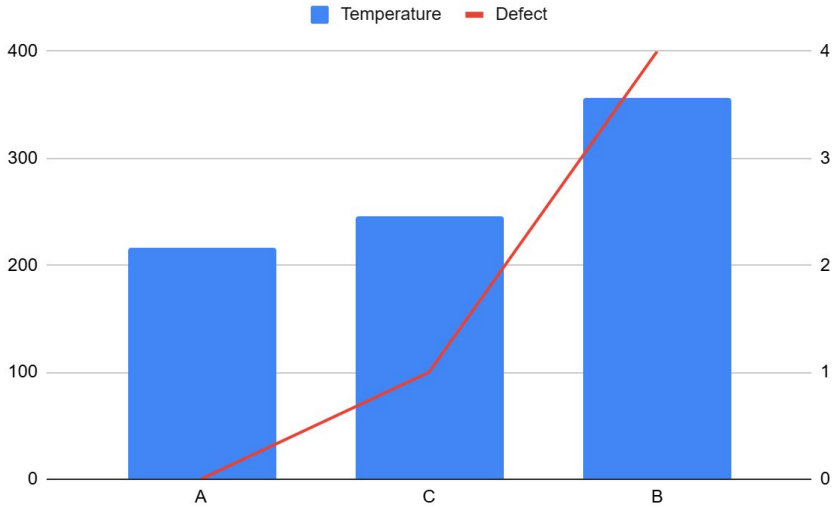
Check Alternative Plots



X-axis is temperature
Plant B has high temperature

Clustering can also be used to derive such description.

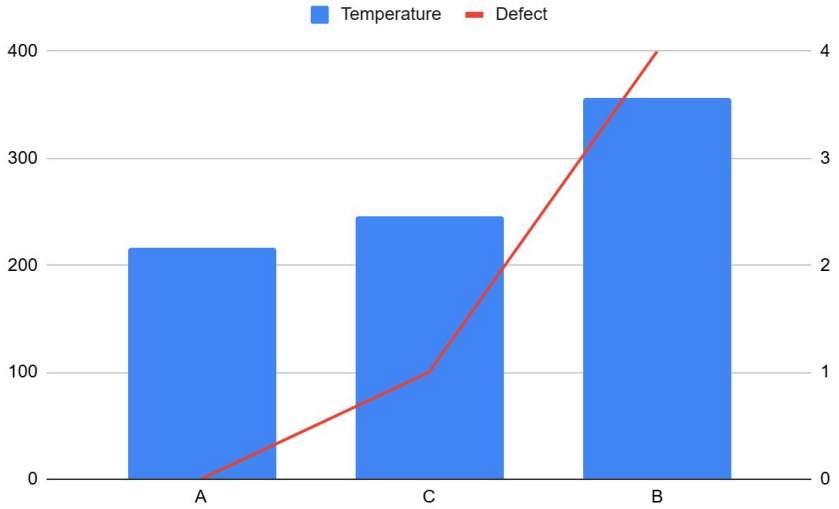
Correlation is not Causation



Temperature and defect are correlated.

But we can not conclude that high temperature is causing defects.

Design of Experiments for Causation



Temperature and defect are correlated.

But we can not conclude that high temperature is causing defects.

We need to perform Design of Experiments to find out cause.

DoE for Causation

First we check why temperature is high for plant B

Is there a problem with the cooling machine ?

If we fix the cooling issue then we perform trials with

Temperature high

Temperature low

Record the observations and perform significance test or ANOVA whichever applicable to nullify random effects.

Predictive Analytics

Suppose, temperature is identified as the cause of the defect.

What should be the correct level of temperature ?

Can we estimate that ?

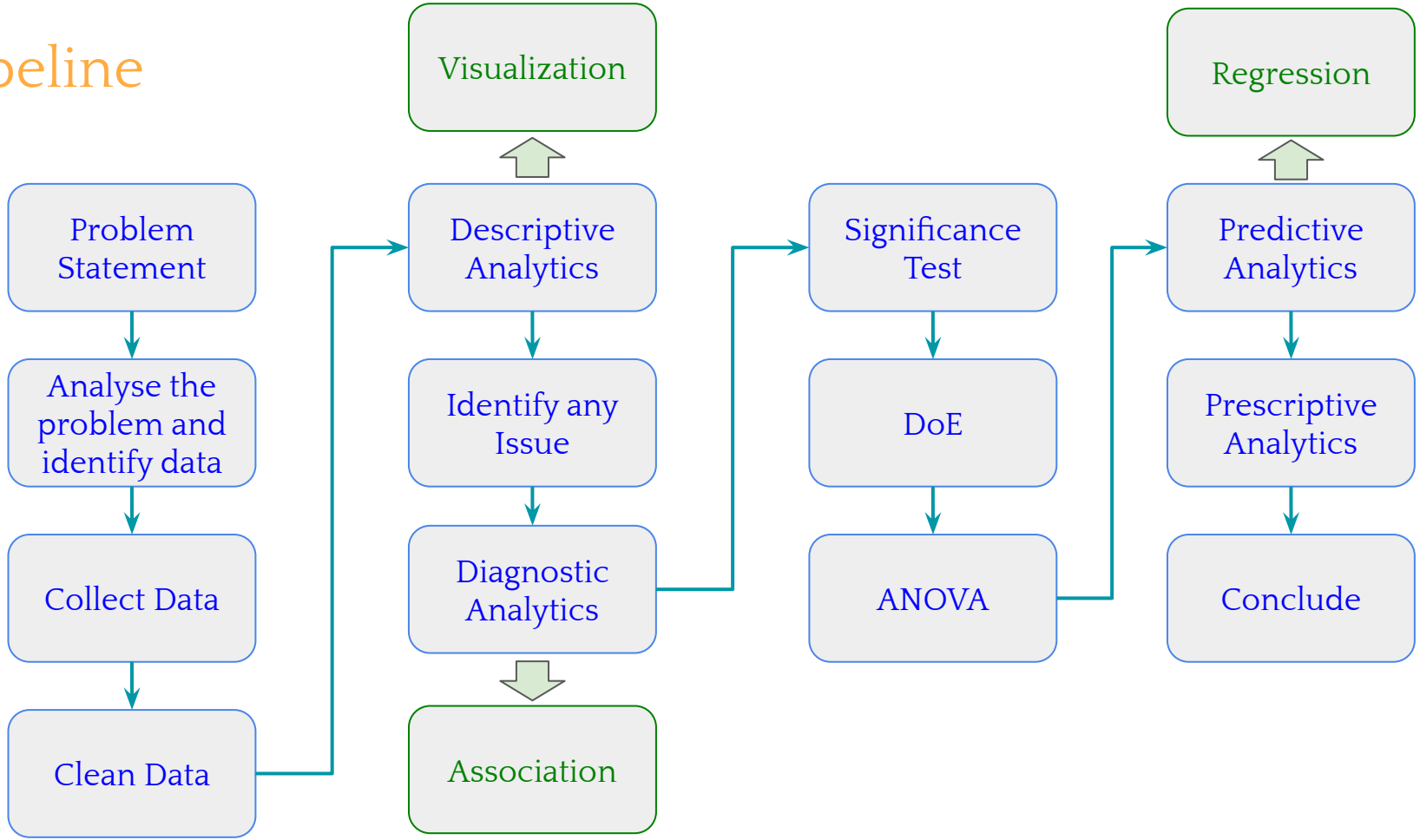
We use predictive analytics to do that.

Prescriptive Analytics

Finally the Data Analytics team of Indian Chemicals can submit a report to the management with their analysis and recommendations.

Based on the data and analysis, we prescribe that the temperature of a plant should be restricted below 80.

Pipeline


Problem Statement
Analyse the problem and identify data
Collect Data
Clean Data
Descriptive Analytics
Identify any Issue
Diagnostic Analytics
Visualization
Association
Significance Test
DoE
ANOVA
Predictive Analytics
Prescriptive Analytics
Conclude
Regression

Keep in Touch

https://nmrinl.github.io/dataanalytics.html

https://nmrinl.github.io/

mrinal@iitpkd.ac.in