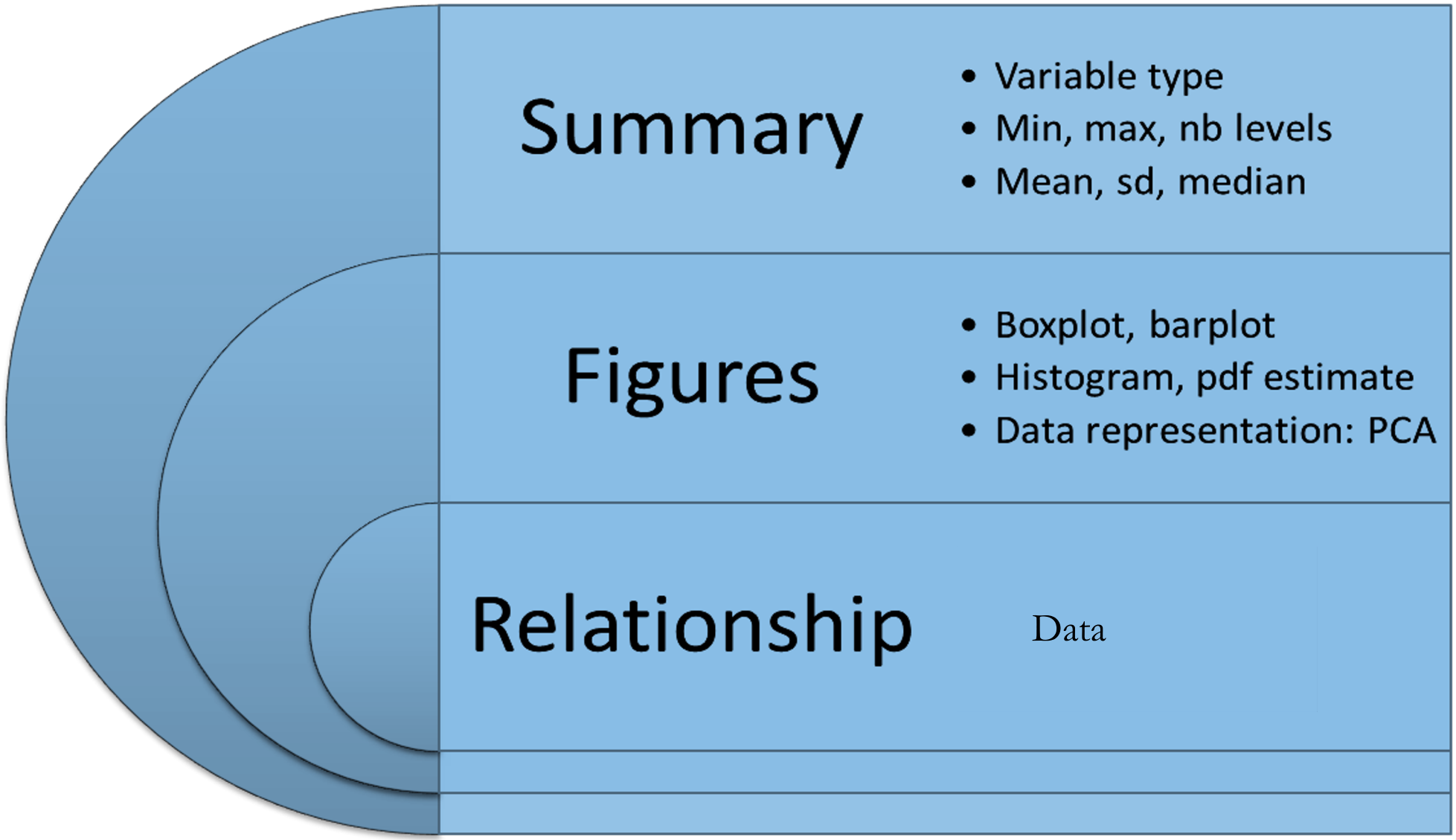
Data Analysis Concepts



Data

Data is a set of values of qualitative or quantitative variables.

Quantitative – Numbers, tests, counting, measuring
Qualitative – Words, images, observations, conversations, Video

Methods of Data Collection

There are several methods of collecting data,

- Surveys
- Observation
- Interview
- Questionnaire

- Manual Data Entry
- IoT
- Camera
- Web scrapping

Methods of data Collection

- Manual Data Entry
- IoT
- Camera
- Web scrapping
- Speech

IoT Simulation

Tweets

Need for Data Analysis

Data Analysis– Life & Death

- In 1987 a disaster was caused by many factors and one of the important factors were poor data Analysis…

- http://www.history.com/topics/challenger-disaster

Data Analysis – Life & Death

- The engineers tried so hard to tell NASA that there was a problem…But NASA management still ahead with launch .

- why did they still launch?

Challenger Disaster

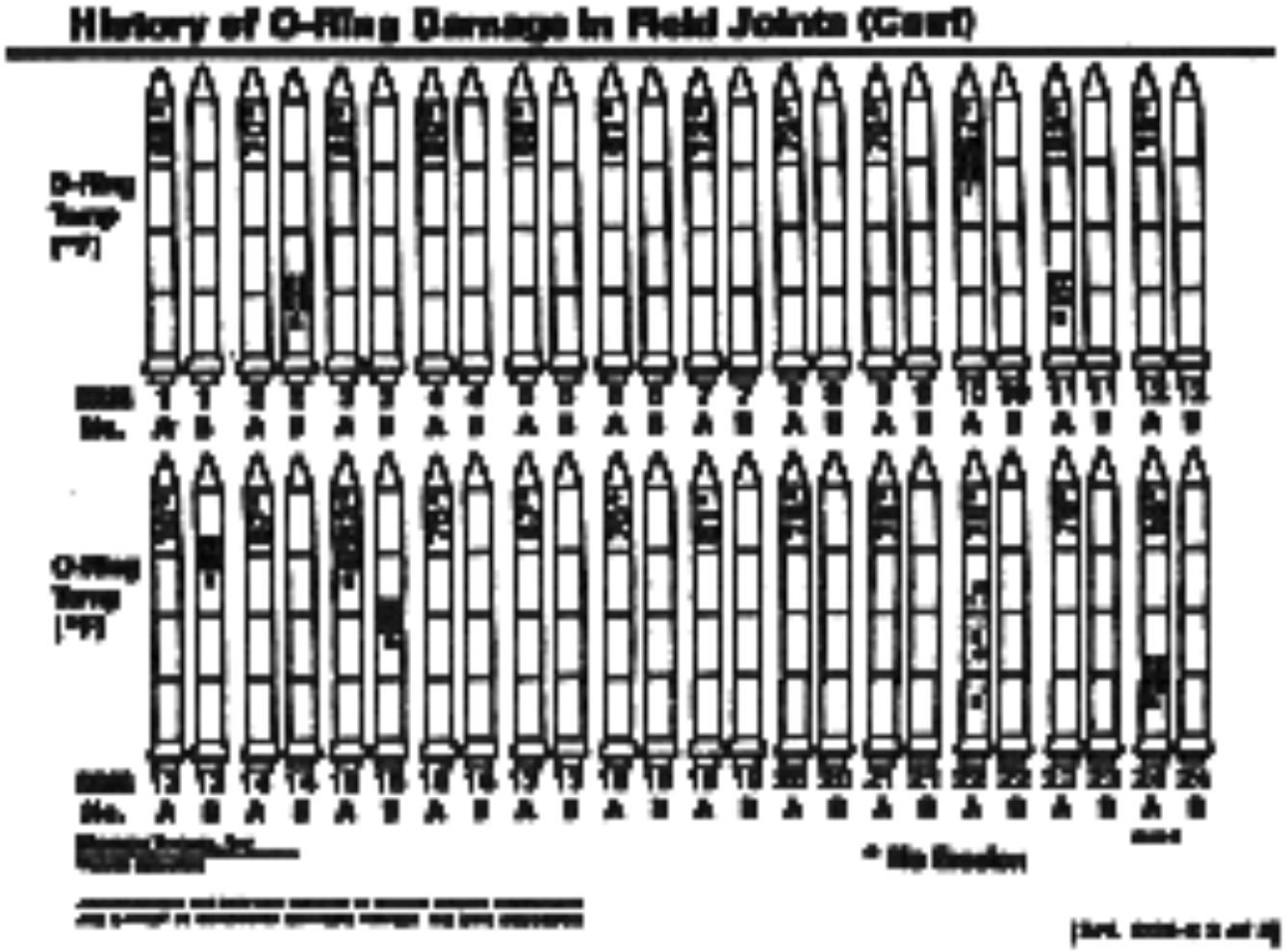
- There is a great deal of evidence that poor data analysis led to poor decision making on the day of the Challenger explosion.

What was the problem?

- The Challenger exploded due to the cold-induced failure of the O-ring seals.
- Was this unexpected? No.
  - Morton Thiokol and team recommended against the launch.
  - Management overruled them.

But…

- The first 24 launches were successful…
- **The clear proximate cause: An inability to convincingly draw a link between temperatures and O-ring failures.**



Data

Order of Flight	Ambient Temperature	Number of failures
1	66	0
2	70	1
3	69	0
4	80	0
5	68	0
6	67	0
7	72	0
8	73	0
9	70	0
10	57	1
11	63	1
12	78	0
13	70	1
14	67	0
15	53	3
16	75	0
17	67	0
18	70	0
19	81	0
20	76	0
21	79	0
22	75	2
23	76	0
24	58	1

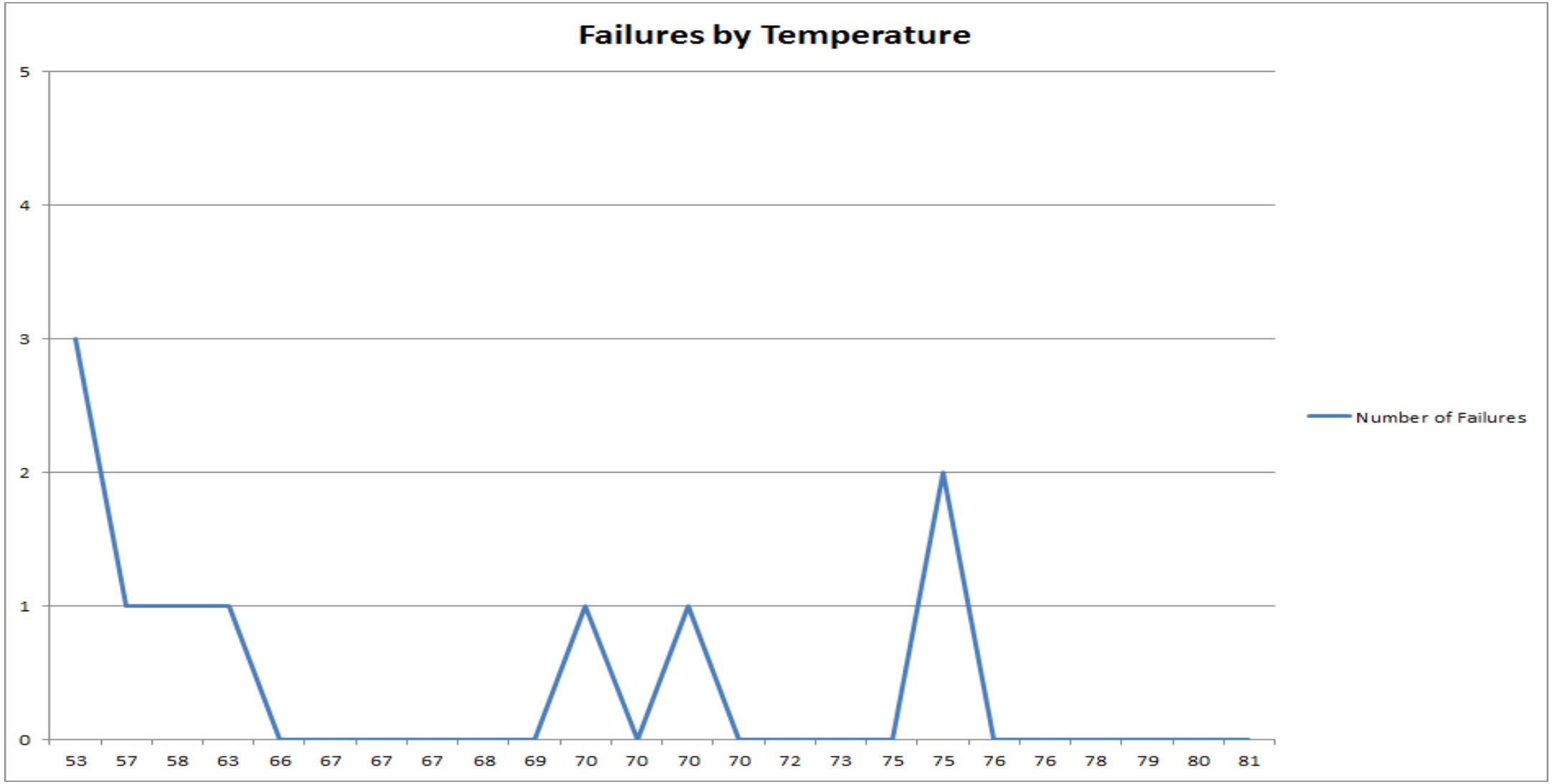
- What insight do you gain from this table?
- Are there issues?
- Can you make a decision?

Better?

Ambient Temperature	Number of failures
53	3
57	1
58	1
63	1
66	0
67	0
67	0
67	0
68	0
69	0
70	1
70	0
70	1
70	0
72	0
73	0
75	0
75	2
76	0
76	0
78	0
79	0
80	0
81	0

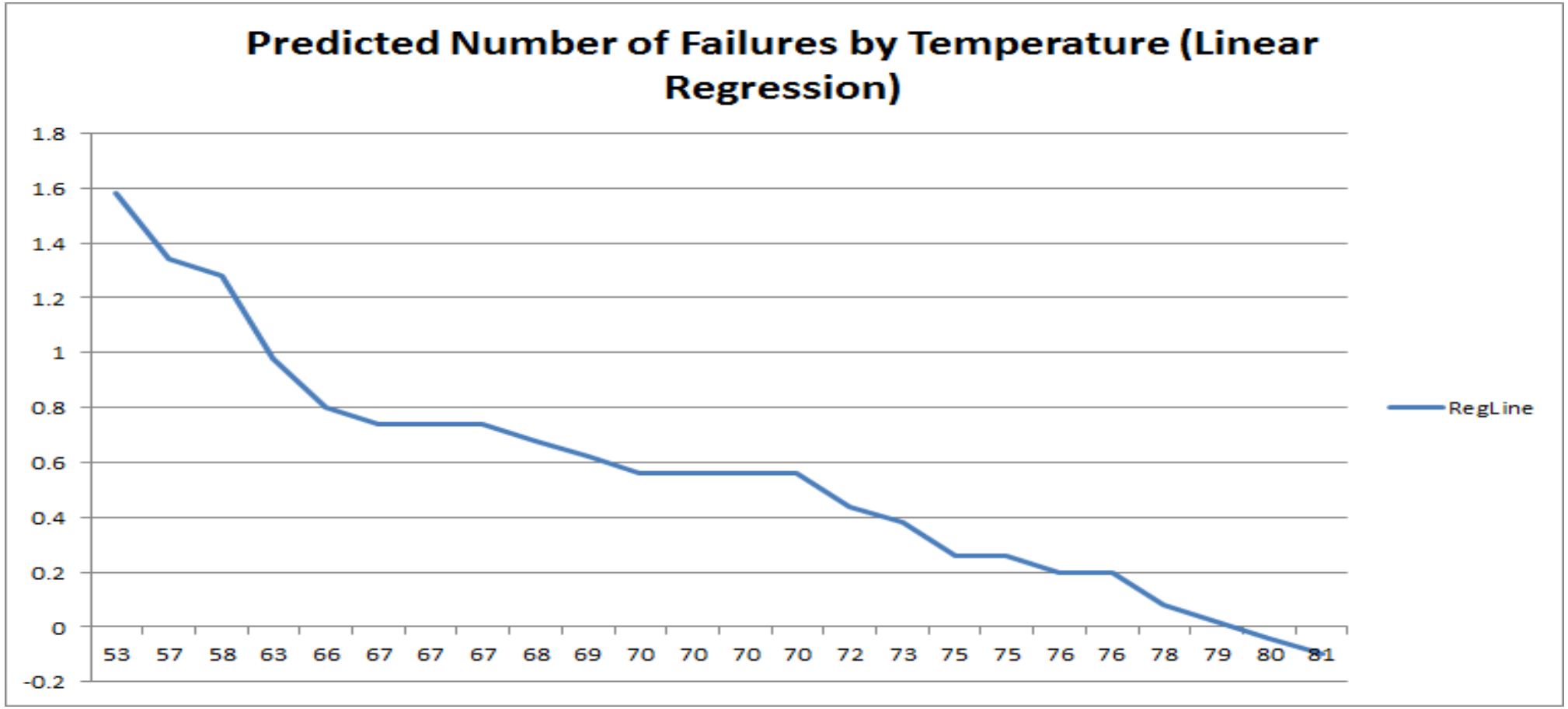
- Can we make better sense of this?
- Can you make a decision?

How about something VISUAL?



How about something predicting failures using analytics ?

- Are there still issues with this?



Let's go back to data again

Data Nuances

- **NOISY DATA** – value may not be accurate
  - Sensor Malfunctioning, Sensor Biased, Sensor Resolution,…
  - Call center notes, Transcription errors, Data entry errors
  - Comments → Tweets →Blogs → News → Scientific Papers
- **MISSING FEATURES** – some feature values might be missing
  - Sensor went down, Communication/Storage failure, Human error
- **NON-NORMAL FEATURE DISTRIBUTIONS**
  - Exponential, log-normal distributions are more common than normal
  - Taking log of features helps
- **HETEREGONEOUS FEATURES** - ranges, scales, distributions.
  - E.g. Age, Income, Temperature, RBC Counts, Blood Pressure,…
- **MULTI-MODALITY FEATURES**
  - Mix of numeric, symbolic, series, text, and image PER data point!

Data Analysis

1) **Describe a dataset**: Number of rows/columns, missing data, data types, preview.

2) **Clean data :** Handle missing data, invalid data types, incorrect values and outliers

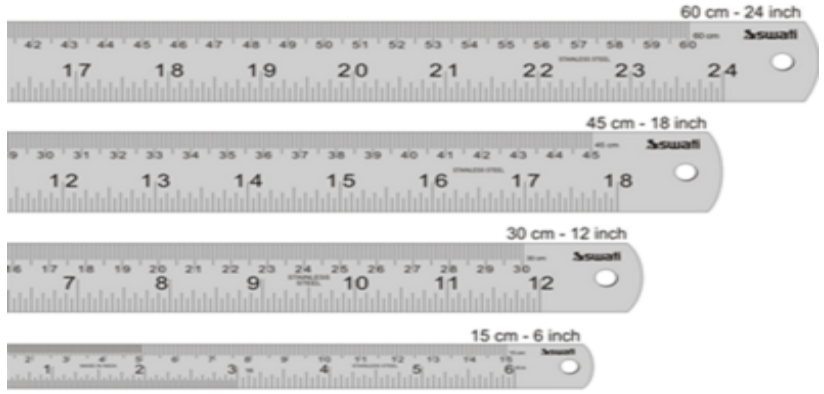
3) **Visualize data distributions:** Bar charts, histograms, box plots.

4) **Calculate and visualize: C**orrelations (relationships) between variables, Heat map

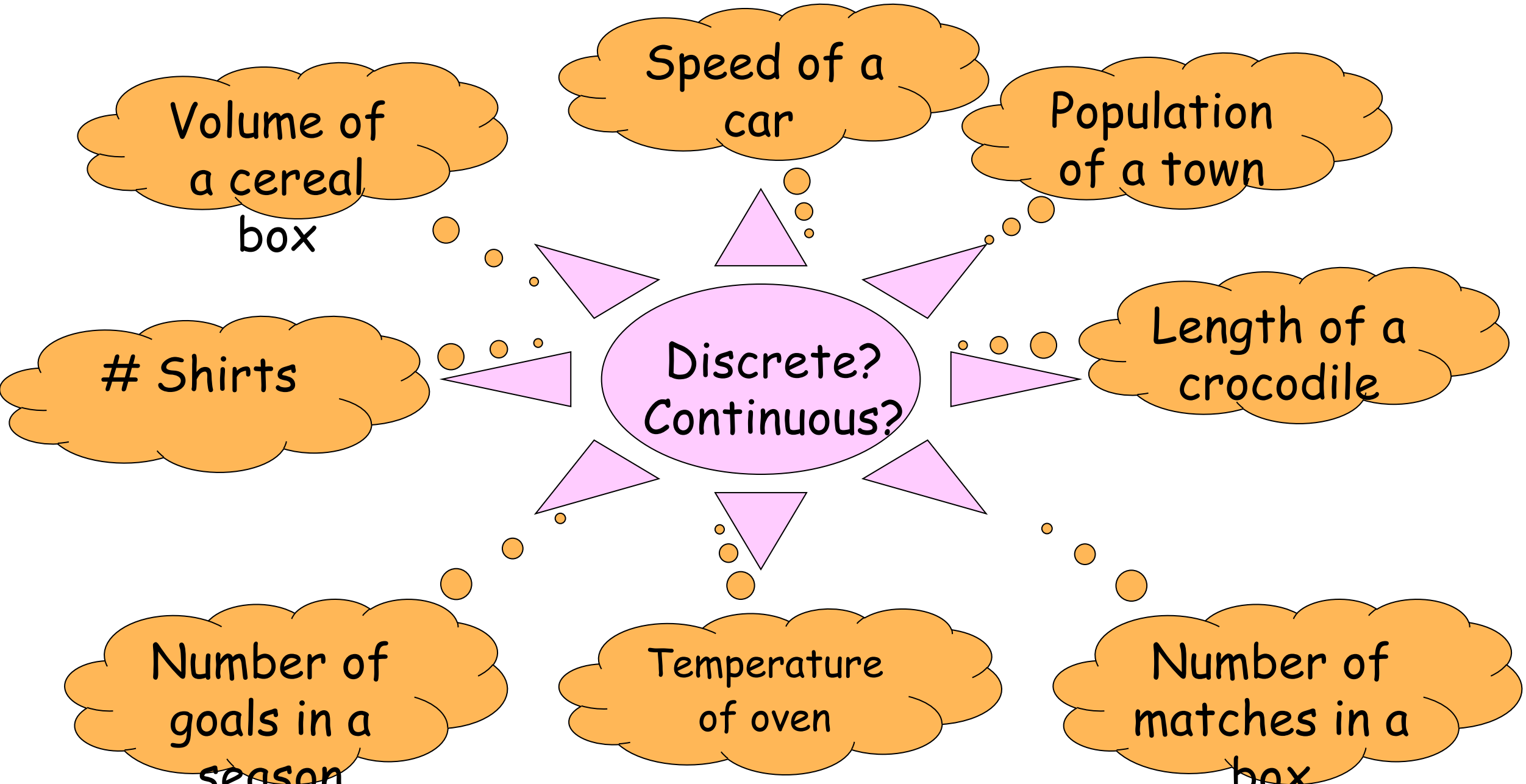
Describe Data Set – Basic Statistics

1. Data Types – Continuous, Discrete, Nominal, Ordinal, Interval, Ratio,
2. First, second, third & fourth moment business decisions
3. Graphical representation – Barplot, Histogram, Boxplot, Pareto chart, Scatter diagram, Correlation

Data Types – Continuous & Discrete



Group the following as either discrete or continuous data.



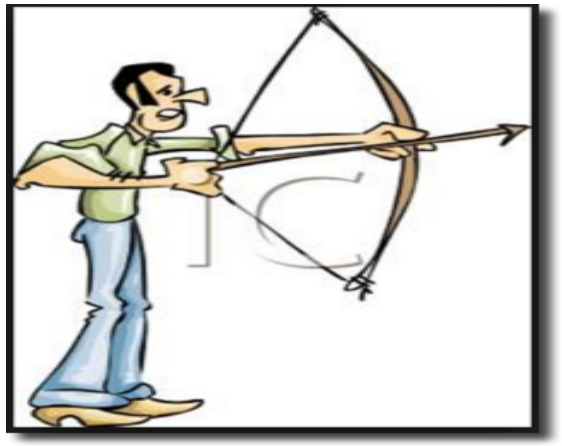
Data come in many flavors ...

Type of data	Definition	Example
Nominal	Categories	Your previous degree
Ordinal	Can be ranked / ordered but not measured	Business school rankings
Interval scale	Intervals are meaningful but not ratios	Temperature in Fahrenheit or Celsius
Ratio scale	Ratios are meaningful	Sales of a new product

Source of data	Definition	Example
Observational	Analyst does not control data generating process	Stock returns on BSE
Experimental	Analyst has good control over data generation	Drug efficacy in clinical trials

Measures of Central Tendency

Central Tendency	Population	Sample
Mean / Average	$\mu = \frac{\Sigma(x_i)}{N}$	$\bar{x} = \frac{\Sigma(x_i)}{n}$
Median	Middle value of the data	
Mode	Most occurring value in the data	



"Every American should have above average income, and my Administration is going to see they get it." – American President

Sample Mean for a Distribution

For a discrete function

Σ y means, "Add up all the Y's"

$$\bar{X} = \hat{\mu} = \sum_{i=1}^{N} x_i / N = \frac{X_1 + X_2 + \ldots X_N}{N}$$

Examples:

Coating weights: 8.47, 8.67, 9.34, 7.99

$$\text{Coating AVERAGE} = \frac{8.47 + 8.67 + 9.34 + 7.99}{4} = 8.62$$

Mean = Average

Sample Median

Assume that x1, x2, …xn is a list of sample data sorted in ascending order. Then…

$$\tilde{X} = \begin{cases} \text{middle value, if } n \text{ is odd} \\ \text{the average of the two middle values, if } n \text{ is even} \end{cases}$$

Mode

The modal value of a set of data is the most frequently occurring value.

Find the mode for: 2, 6, 3, 9, 5, 6, 2, 6

It can be seen that the most frequently occurring value is 6. (There are 3 of these)

Bi model and Multi model

Mode for:

1) 1,2,3,3,3,4,4,4,5,6,7
2) 2,2,3,10,11,17,3,10

Measures of Variability

The mean, mode, and median do a nice job in telling where the center of the data set is, but often we are interested in more…

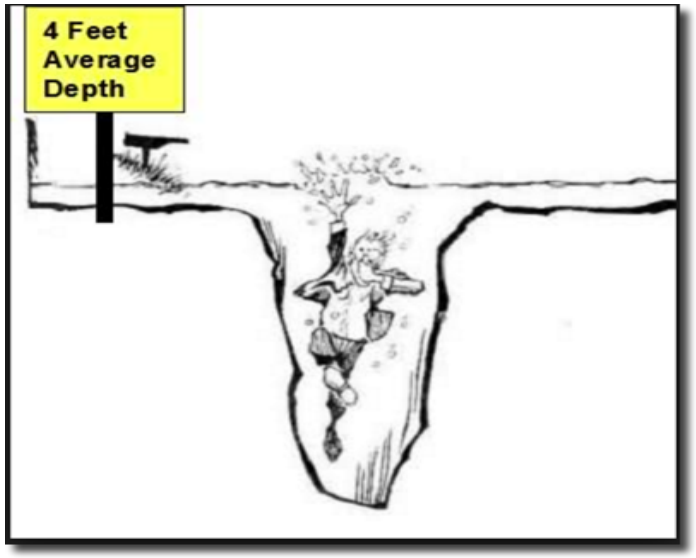
For example, a pharmaceutical engineer develops a new drug that regulates sugar in the blood.
Suppose she finds out that the average sugar content after taking the medication is the optimal level.

This does not mean that the drug is effective. There is a possibility that half of the patients have dangerously low sugar content while the other half has dangerously high content. Instead of the drug being an effective regulator, it is a deadly poison.

What the pharmacist needs is a measure of how far the data is spread apart. This is what the variance and standard deviation do

Measures of Dispersion

Dispersion	Population	Sample
Variance	$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$	$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$
Standard Deviation	$= \sqrt{\frac{\Sigma (X_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$
Range	Max – Min	



We define the *variance* to be

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$$

standard deviation to be

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2}$$

Range

The "Range" for a data set is the difference between the largest value and smallest value contained in the data set. First reorder the data set from smallest to largest then subtract the first element from the last element

Data Set = 2, 5, 9, 3, 5, 4, 7
Reordered = 2, 3, 4, 5, 5, 7, 9

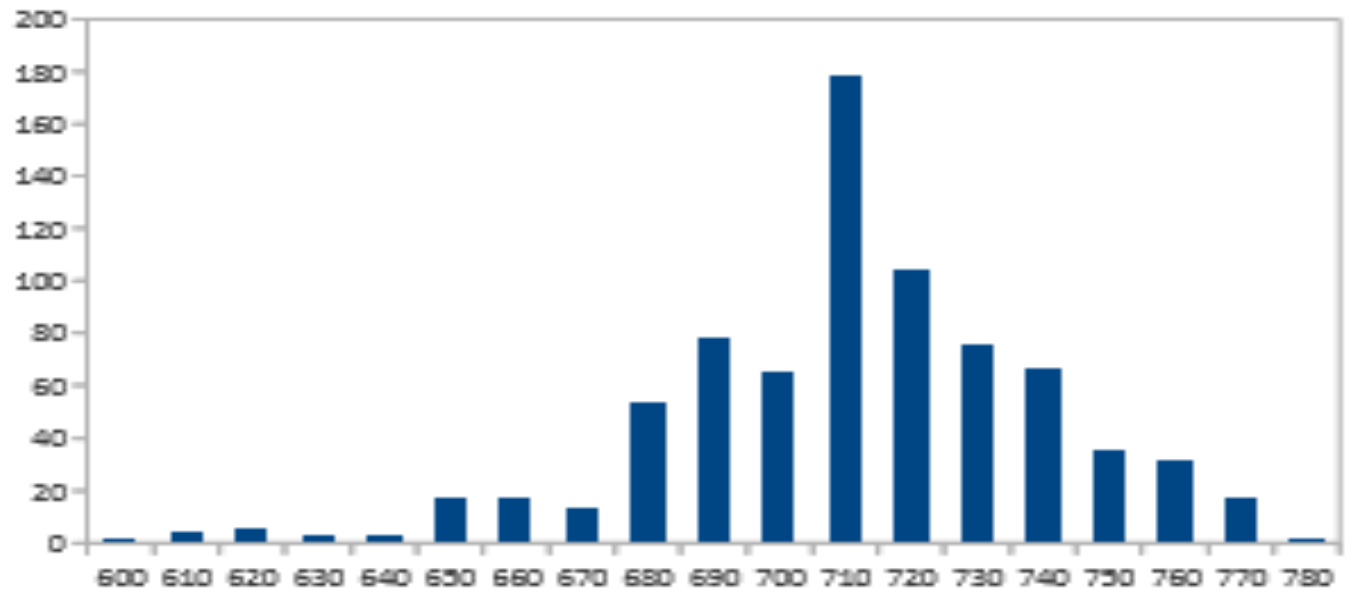
Range = (9 - 2) = 7

Visualization Techniques

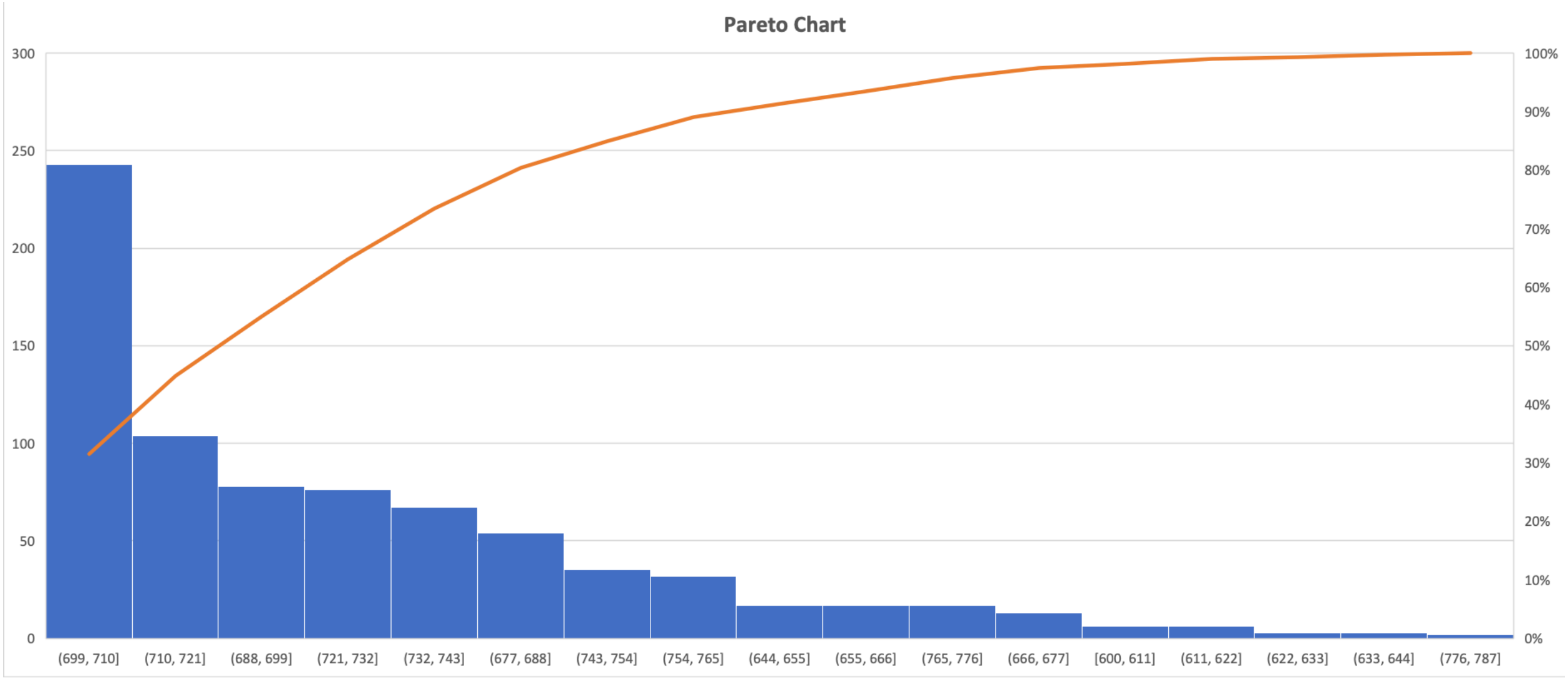
GMAT Scores of an MBA Class

610	730	590	610	.	.	.	650	630
640	630	540	660	.	.	.	610	540
690	610	520	640	.	.	.	720	650
610	650	660	530	.	.	.	600	730
710	600	760	690	.	.	.	500	720
610	650	660	710	.	.	.	450	600
630	610	680	730	.	.	.	700	690
530	550	730	690	.	.	.	670	540
630	720	610	710	.	.	.	600	600
690	600	730	540	.	.	.	560	770

Pictorial summary of data: A bar chart



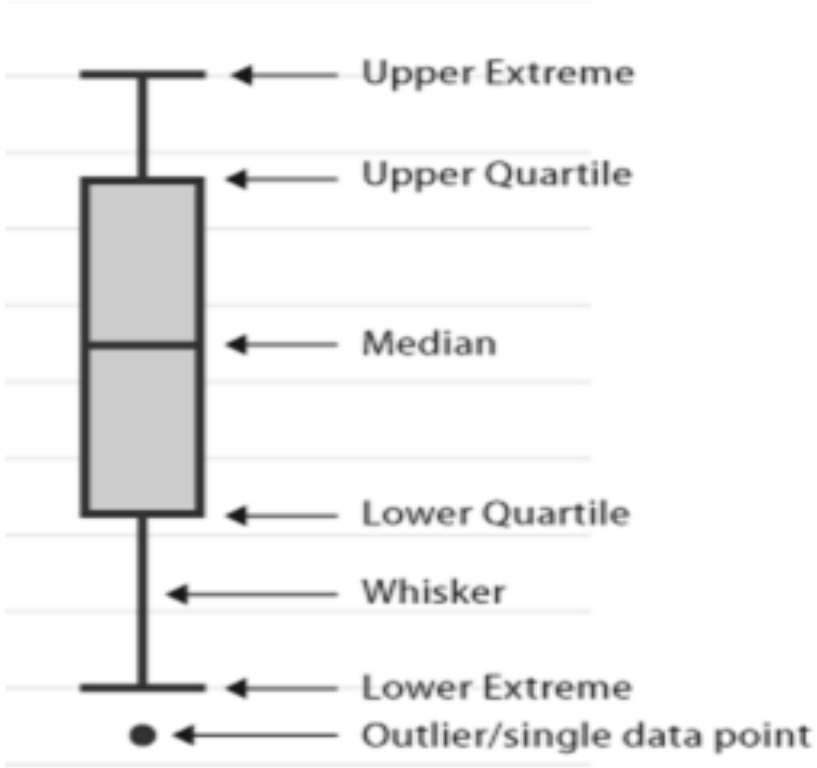
Pareto Chart (80 /20 rule)



Graphical Techniques – Box Plot

Range(IQR): The middle half of a data set falls within the inter- quartile range

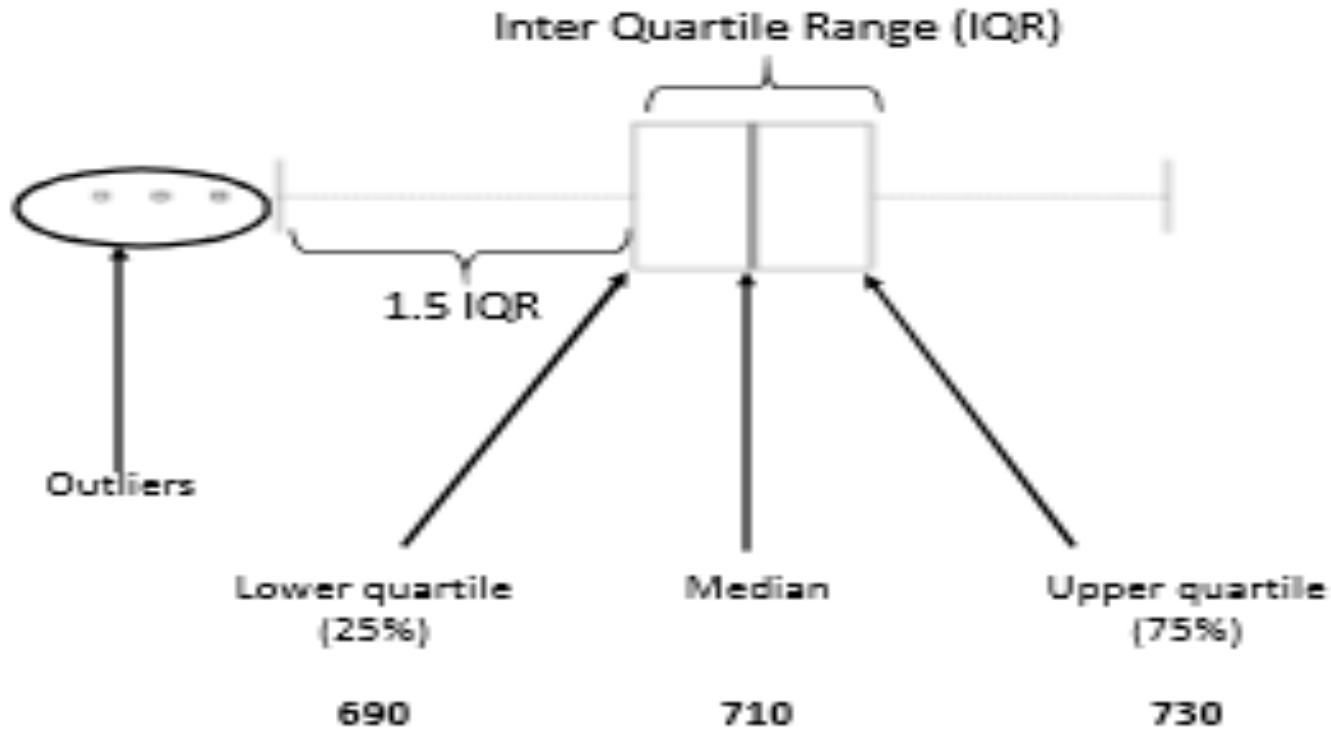
Inter-quartile



Box Plot : This graph shows the distribution of data by dividing the data into four groups with the same number of data points in each group. The box contains the middle 50% of the data points and each of the two whiskers contain 25% of the data points. It displays two common measures of the variability or spread in a data set

Range : It is represented on a box plot by the distance between the smallest value and the largest value, including any outliers. If you ignore outliers, the range is illustrated by the distance between the opposite ends of the whiskers

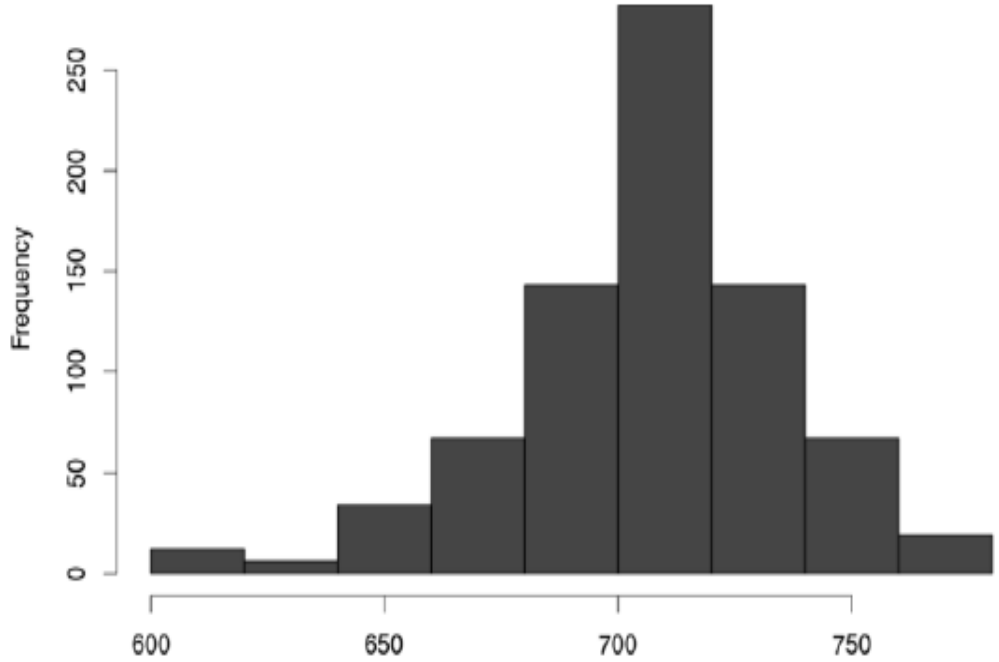
Boxplot

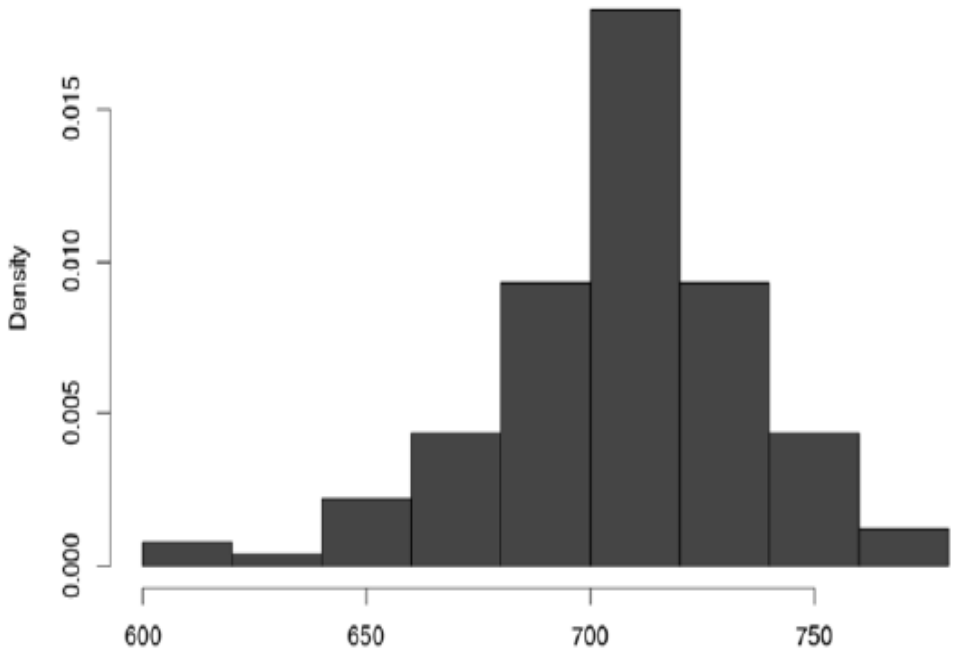


A boxplot displays the prominent quartiles of the data along with outliers

Graphical Techniques – Histogram

A Histogram Represents the frequency distribution, i.e., how many observations take the value within a certain interval.





Skewness & Kurtosis

Third and Fourth moments

Skewness

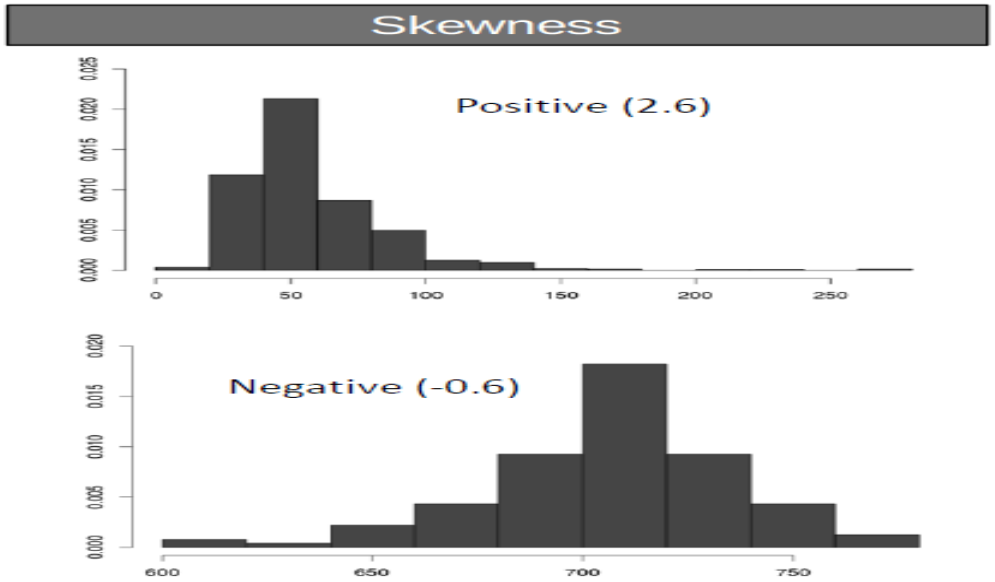
- A measure of asymmetry in the distribution
- Mathematically it is given by

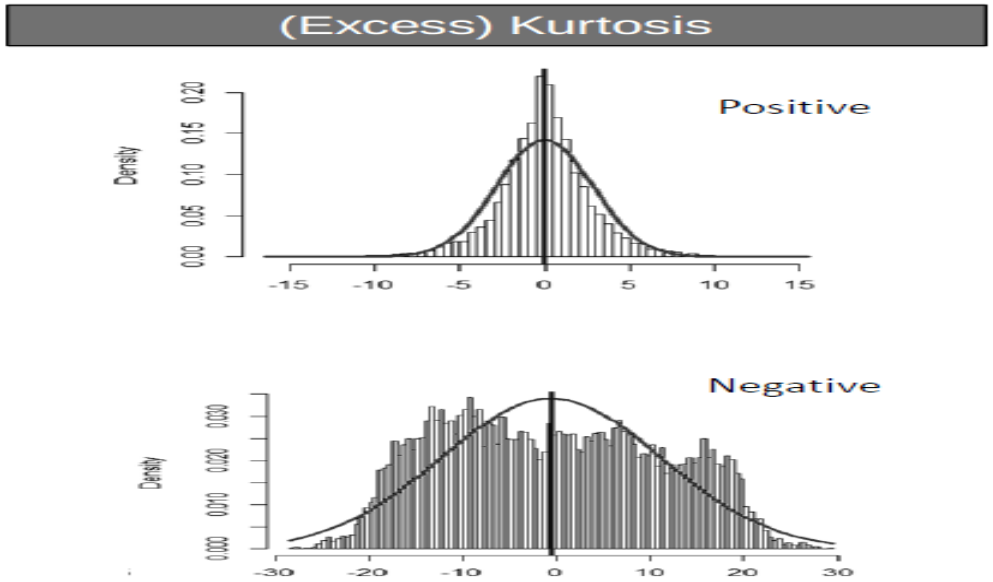
  $E[(x-\mu/\sigma)]^3$
- Negative skewness implies mass of the distribution is concentrated on the right

Kurtosis

- A measure of the "Peakedness" of the distribution
- Mathematically it is given by

  $E[(x-\mu/\sigma)]^4 - 3$
- For Symmetric distributions, negative kurtosis implies wider peak and thinner tails



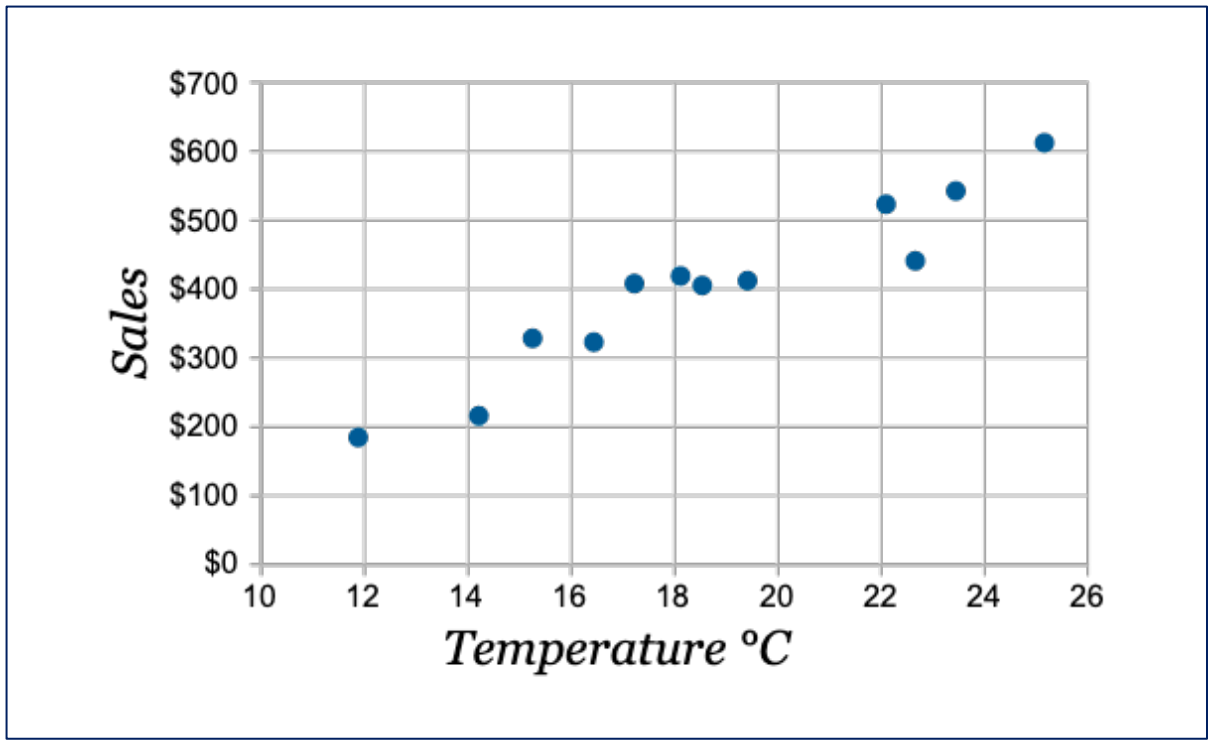


Scatter Plot

Scatter Plot

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. Scatter plots are used to observe relationships between variables.

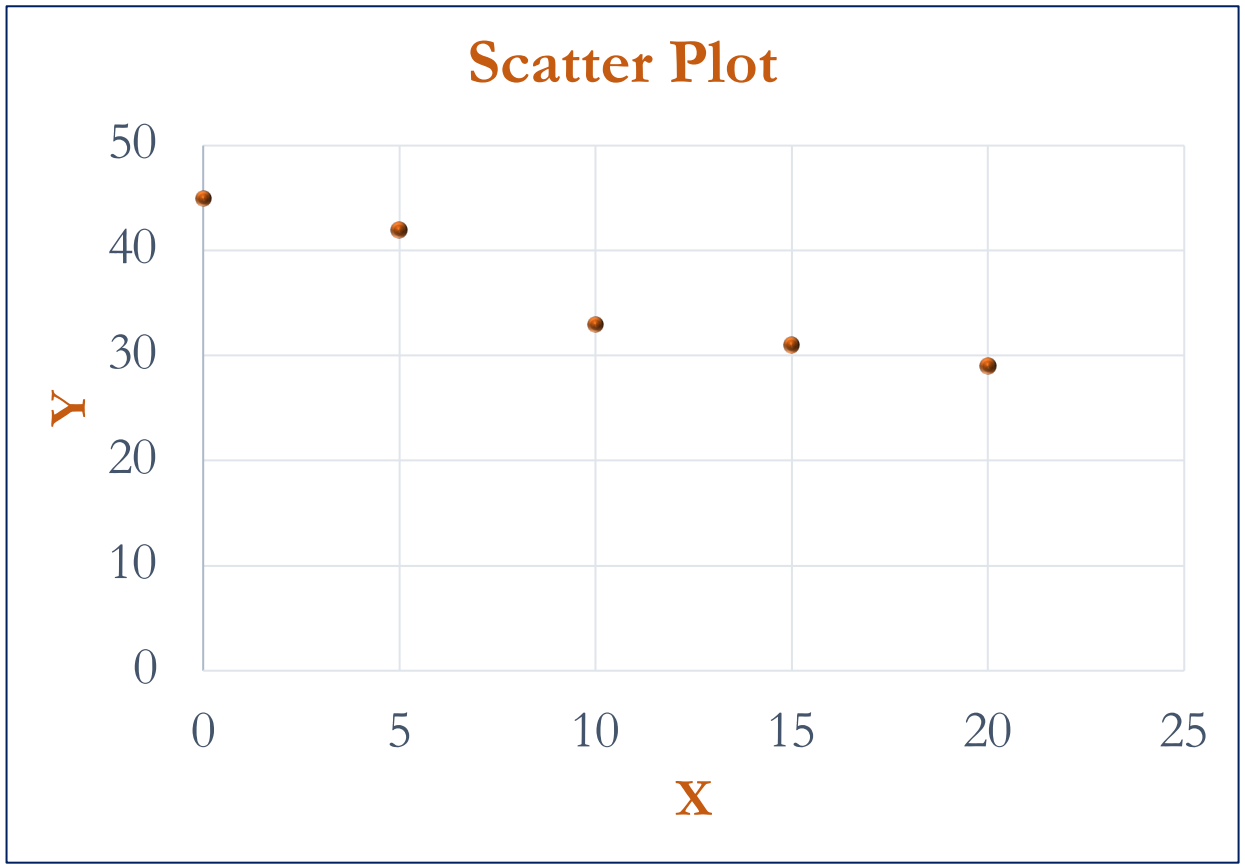
Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



Which variable affects which one?

Scatter Plot

Cigarettes (X) in Years	Lung Capacity (Y)
0	45
5	42
10	33
15	31
20	29



Correlation

Pearson Correlation

Correlation is a bi-variate analysis that measures the strength of linear association between two variables and the direction of the relationship. Correlation is a statistical technique used to determine the degree to which two variables are linearly related.

[-1 ..... 0 ..... +1]

High correlation

Low correlation

High correlation

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{(Y-\bar{Y})^2}}$$

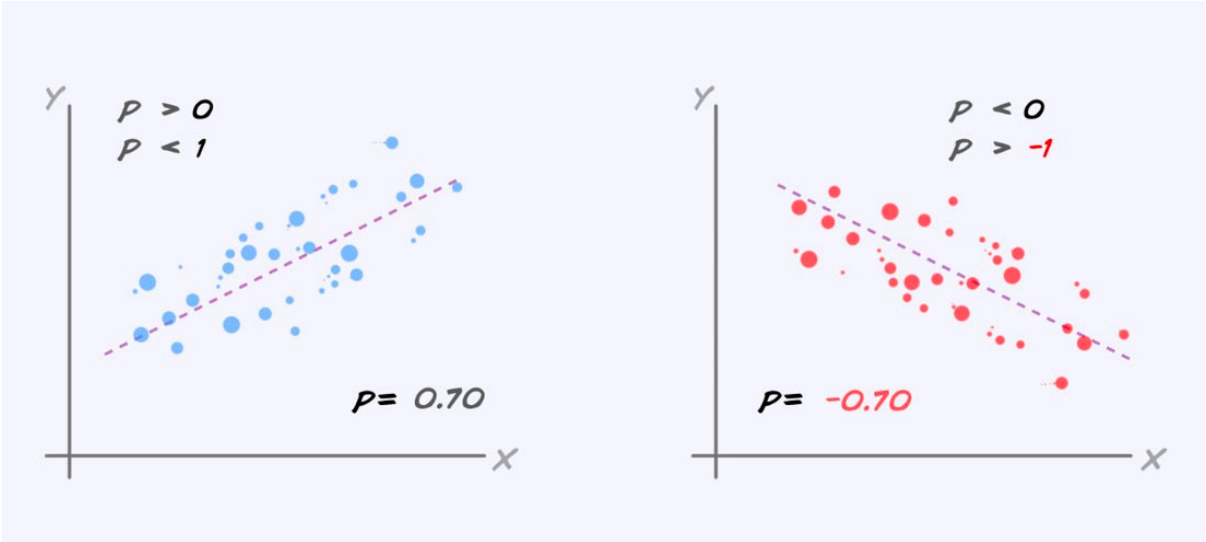
Where, $\bar{X}$ = mean of X variable

$\bar{Y}$ = mean of Y variable

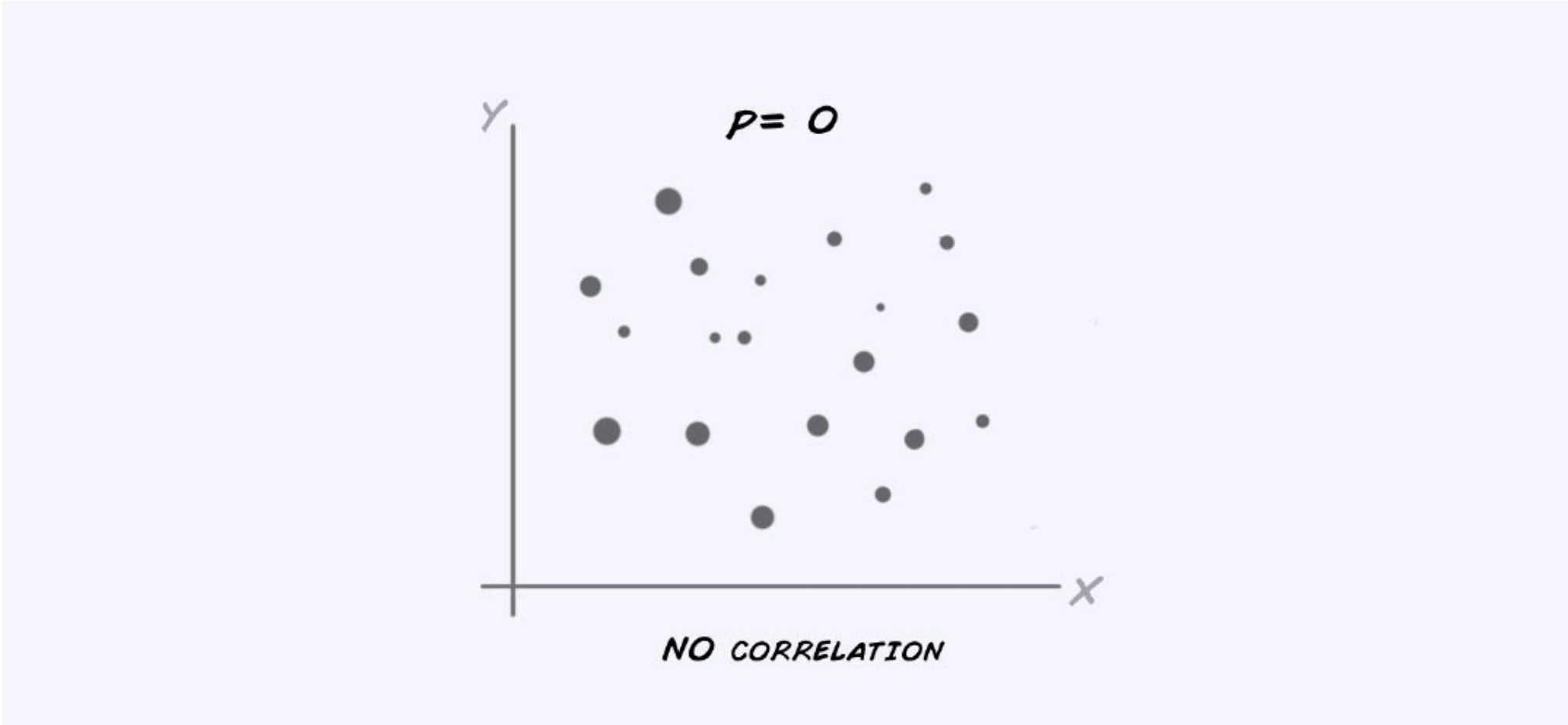
Correlation r - Interpretation

- Positive r indicates positive linear association between x and y or variables, and negative r indicates negative linear relationship
- r – always between -1 and +1
- The strength increases as r moves away from zero toward wither -1 or +1
- The extreme values +1 and -1 indicate perfect linear relationship (points lie exactly along a straight line)
- Graded interpretation : r 0.1-0.3 = weak; 0.4-0.7 = moderate and 0.8-1.0=strong correlation

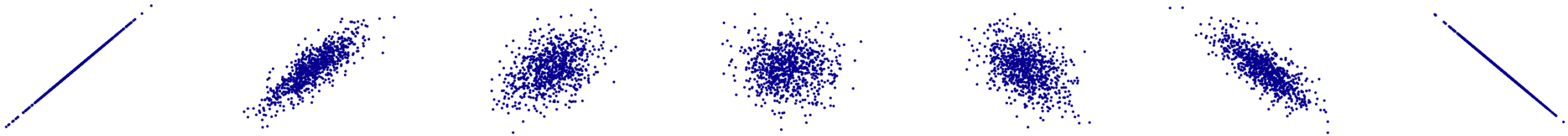
Correlation



Correlation



1 0.8 0.4 0 -0.4 -0.8 -1



1 1 1 -1 -1 -1

0 0 0 0 0 0 0

Source: https://en.wikipedia.org/wiki/Correlation_and_dependence

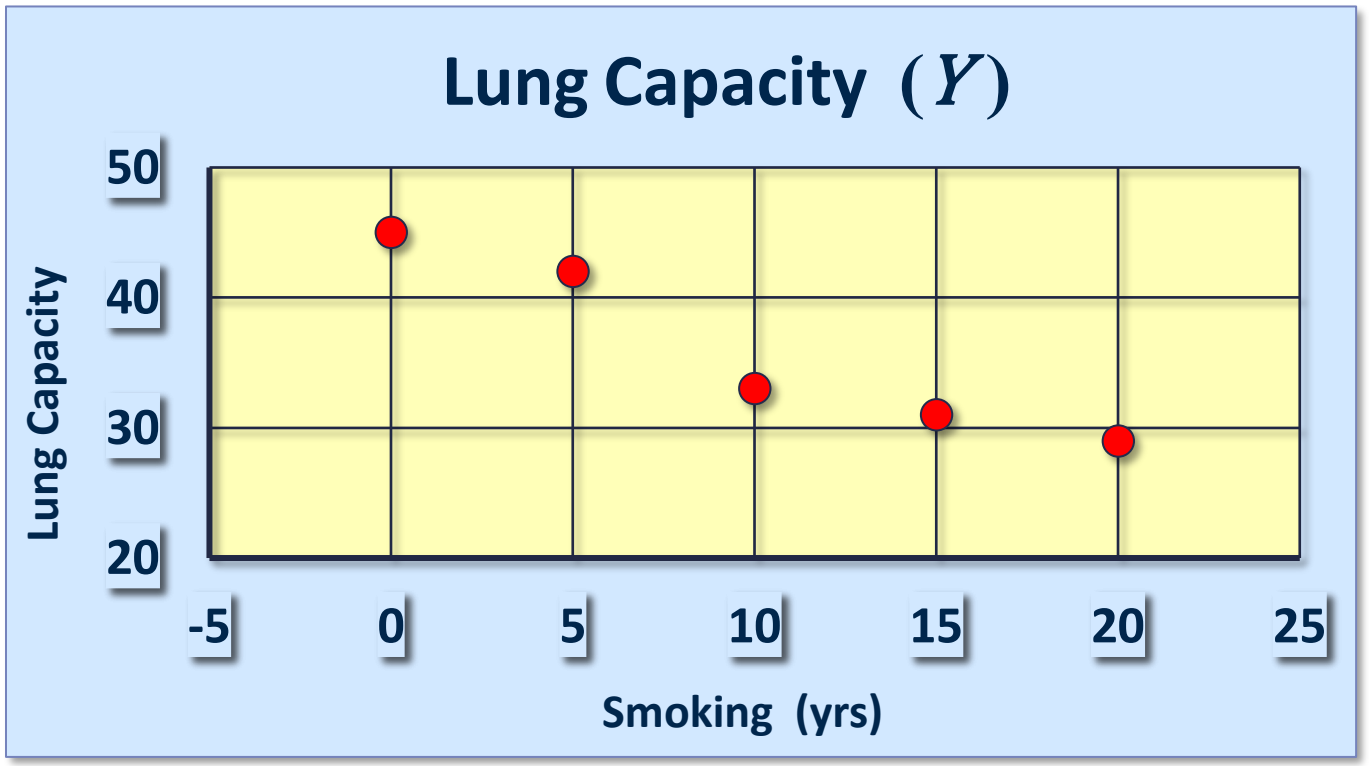
Refer for non linear association : https://towardsdatascience.com/rip-correlation-introducing-the-predictive-power-score-3d90808b9598

Scatter Plot and Correlation : Smoking and Lung Capacity

- Example: investigate relationship between cigarette smoking and lung capacity
- Data: sample group response data on smoking habits, and measured lung capacities, respectively

Smoking v Lung Capacity Data

N	Cigarettes (X)	Lung Capacity (Y)
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29



$$r_{xy} = -0.96$$

- -0.96 implies almost certainty smoker will have diminish lung capacity

Missing Values

What is missing value

Some of the values will be missed in the data set because of various reasons such as human error, machine failures etc

	Ozone	Solar	Wind	Month	Day	Year	Temp	Date
4	NaN	NaN	14.3	5.0	5	2010	56	2010-05-05
5	28.0	NaN	14.9	5.0	6	2010	66	2010-05-06
9	NaN	194.0	8.6	5.0	10	2010	69	2010-05-10
10	7.0	NaN	6.9	5.0	11	2010	74	2010-05-11
23	32.0	92.0	12.0	NaN	24	2010	61	NaT
24	NaN	66.0	16.6	5.0	25	2010	57	2010-05-25
25	NaN	266.0	14.9	5.0	26	2010	58	2010-05-26
26	NaN	NaN	8.0	5.0	27	2010	57	2010-05-27
31	NaN	286.0	8.6	6.0	1	2010	78	2010-06-01

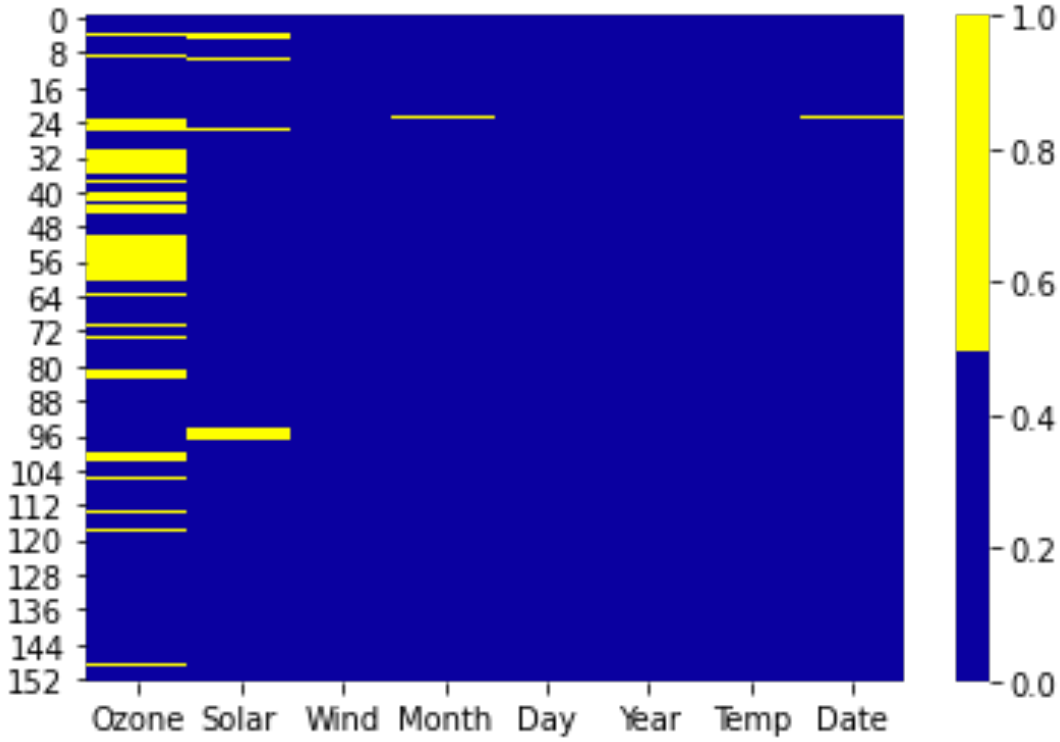
Missing Values: How to find?

Missing Data Heatmap:

When there is a smaller number of features, we can visualize the missing data via heatmap.

	Ozone	Solar	Wind	Month	Day	Year	Temp	Date
4	NaN	NaN	14.3	5.0	5	2010	56	2010-05-05
5	28.0	NaN	14.9	5.0	6	2010	66	2010-05-06
9	NaN	194.0	8.6	5.0	10	2010	69	2010-05-10
10	7.0	NaN	6.9	5.0	11	2010	74	2010-05-11
23	32.0	92.0	12.0	NaN	24	2010	61	NaT
24	NaN	66.0	16.6	5.0	25	2010	57	2010-05-25
25	NaN	266.0	14.9	5.0	26	2010	58	2010-05-26
26	NaN	NaN	8.0	5.0	27	2010	57	2010-05-27
31	NaN	286.0	8.6	6.0	1	2010	78	2010-06-01

The horizontal axis shows the feature name; the vertical axis shows the number of observations/rows; the yellow colour represents the missing data while the blue colour otherwise.



Missing value imputation

Some of the values will be missed in the data set because of various reasons such as human error, machine failures etc

	Ozone	Solar	Wind	Month	Day	Year	Temp	Date
4	NaN	NaN	14.3	5.0	5	2010	56	2010-05-05
5	28.0	NaN	14.9	5.0	6	2010	66	2010-05-06
9	NaN	194.0	8.6	5.0	10	2010	69	2010-05-10
10	7.0	NaN	6.9	5.0	11	2010	74	2010-05-11
23	32.0	92.0	12.0	NaN	24	2010	61	NaT
24	NaN	66.0	16.6	5.0	25	2010	57	2010-05-25
25	NaN	266.0	14.9	5.0	26	2010	58	2010-05-26
26	NaN	NaN	8.0	5.0	27	2010	57	2010-05-27
31	NaN	286.0	8.6	6.0	1	2010	78	2010-06-01

Treatment Methods

- Drop the observation
- Mean Imputation
- Median Imputation
- Statistical (Regression) Imputation