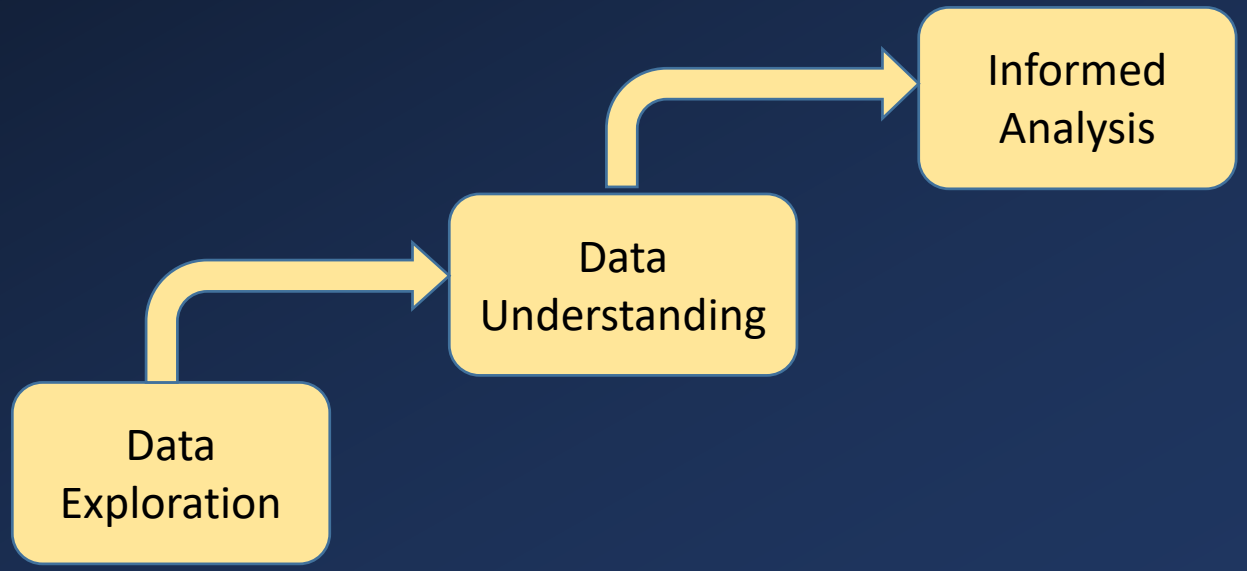
Big Data (MHI222956/MHI225101)

6.2 Data Pre-processing

- A better understanding of the complexity of the data is gained by *exploring your data*.
- Better understanding in turn will guide the rest of the process and lead to more informed analysis.



If you use bad or “dirty” data to train your model, you’ll end up with a bad, improperly trained model that won’t actually be relevant to your analysis.

Good, preprocessed data is even more important than the most powerful algorithms

- Raw, real-world data may not only contain errors and inconsistencies, but it is often incomplete, and doesn't have a regular, uniform design.
- Machines like to process nice and tidy information
- Data pre-processing, also called *Data Preparation*, aims to create data for analysis.
- Data preprocessing is a step that takes raw data and transforms it into a format that can be understood and analyzed by computers.

Data pre-processing

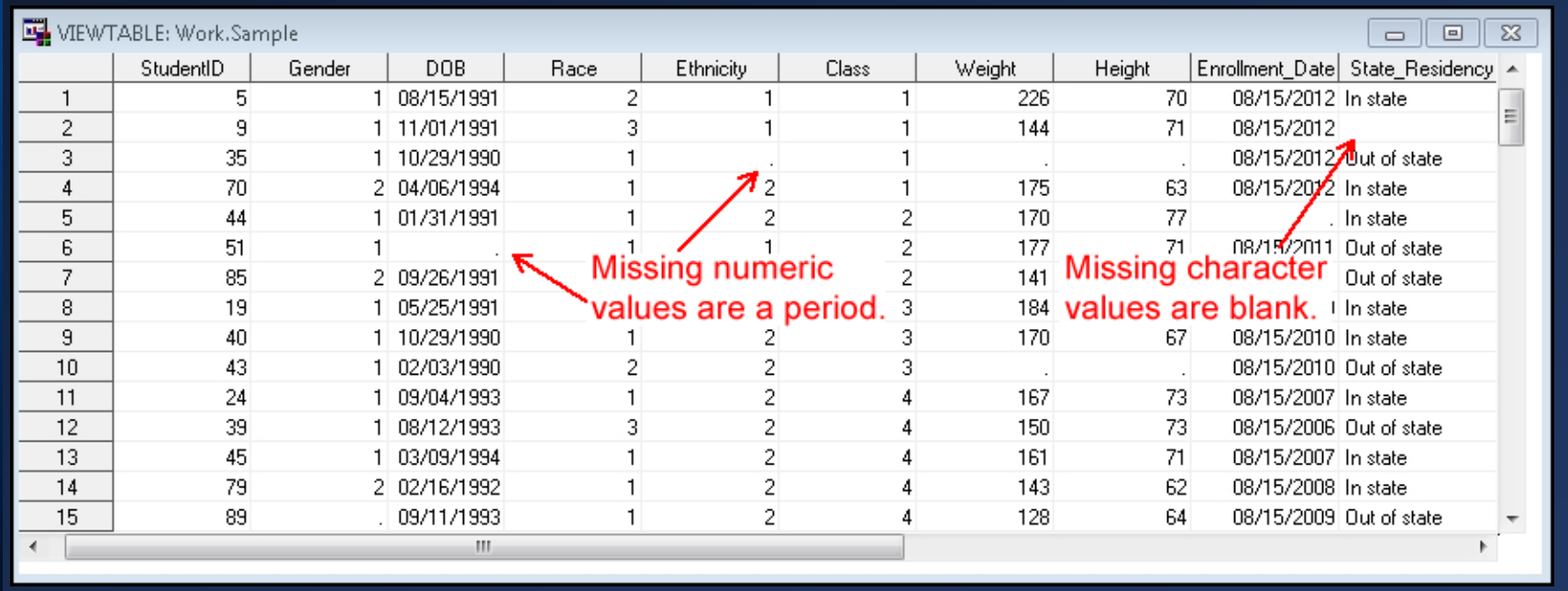
- Data cleaning
  - To clean the data to address data quality issues
  - Missing values
  - Duplicate Data
  - Invalid data
  - Outliers

  Domain knowledge is essential to making informed decisions on how to handle incomplete or incorrect data.

- Data transformation
  - Transform raw data to make it suitable for analysis
  - Scaling
  - Transformation
  - Feature selection
  - Dimensionality reduction
  - Data manipulation

Missing Values

- Occur when no data value is stored for the variable in an observation. (statistics)
- Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.



	StudentID	Gender	DOB	Race	Ethnicity	Class	Weight	Height	Enrollment_Date	State_Residency
1	5	1	08/15/1991	2	1	1	226	70	08/15/2012	In state
2	9	1	11/01/1991	3	1	1	144	71	08/15/2012	
3	35	1	10/29/1990	1	.	1	.	.	08/15/2012	Out of state
4	70	2	04/06/1994	1	2	1	175	63	08/15/2012	In state
5	44	1	01/31/1991	1	2	2	170	77	.	In state
6	51	1	.	1	1	2	177	71	08/15/2011	Out of state
7	85	2	09/26/1991	1	2	2	141	.	08/15/2011	Out of state
8	19	1	05/25/1991	1	2	3	184	71	08/15/2010	In state
9	40	1	10/29/1990	1	2	3	170	67	08/15/2010	In state
10	43	1	02/03/1990	2	2	3	.	.	08/15/2010	Out of state
11	24	1	09/04/1993	1	2	4	167	73	08/15/2007	In state
12	39	1	08/12/1993	3	2	4	150	73	08/15/2006	Out of state
13	45	1	03/09/1994	1	2	4	161	71	08/15/2007	In state
14	79	2	02/16/1992	1	2	4	143	62	08/15/2008	In state
15	89	.	09/11/1993	1	2	4	128	64	08/15/2009	Out of state

- Handling missing values
  - If the number of cases of missing values is extremely small: less than 5% of the sample
    - Drop or omit those values from the analysis
  - If the number of cases of missing values is large
    - In the case of multivariate analysis, it can be better to drop those cases / variables
    - In univariate analysis, imputation can decrease the amount of bias in the data
  - Imputation methods
    - Mean
    - K-nearest neighbors (KNN)
    - etc.

Invalid data

- Categories of invalid data
  - Official missing data values
  - Invalid values that provide no information regarding true value
  - Invalid values that suggest true value

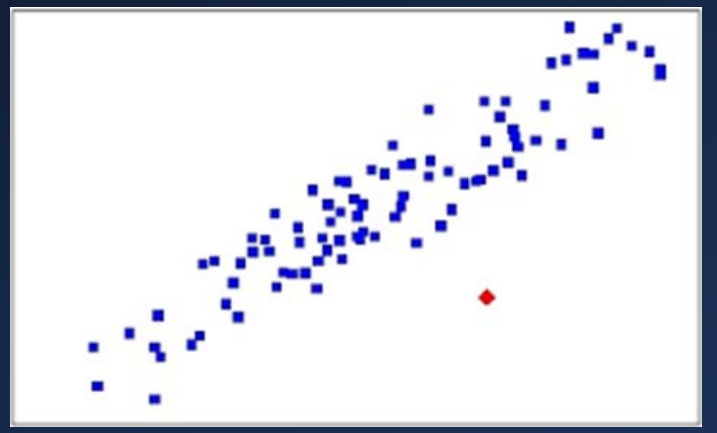
- The problem of invalid data will typically be transformed into one of missing data.

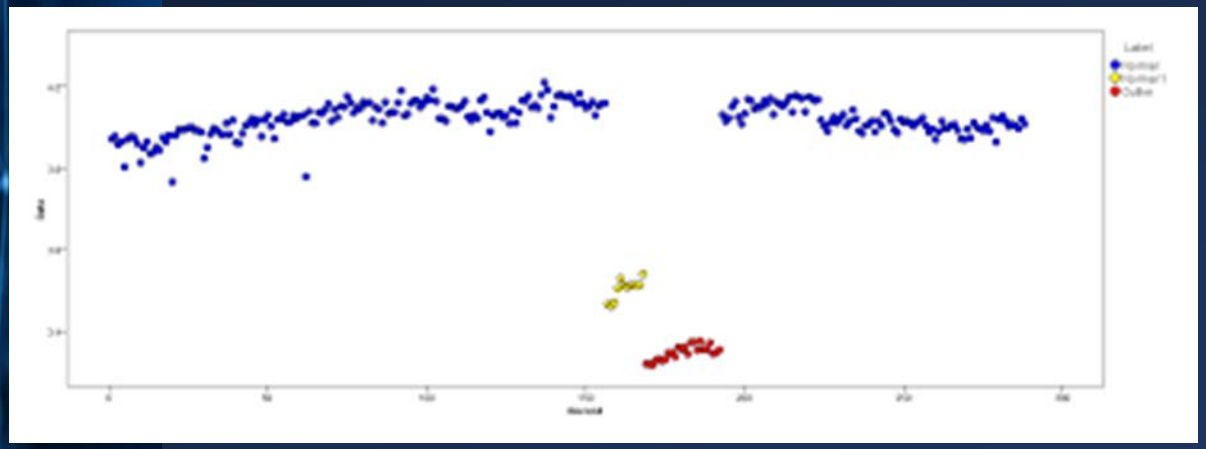
Name	Zip Code
Angela	346412
Sidney	92618
Ratan	8033A
Kiril	11012
Zhou	59285

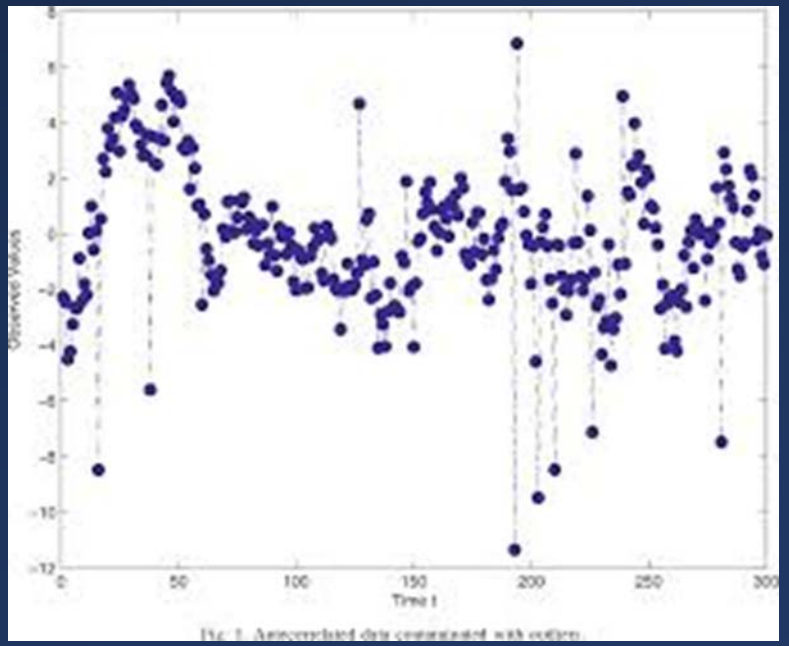
Name	Address
Angela	430 Park Drive
Sidney	780 ★❖©◆ ew Street
Ratan	12443 Mountain Avenue

Outlier

- Identify outlier
- Outlier or Extreme Values?
- Remove outlier
- **Domain Knowledge**







Scaling

- Min-Max scaling
  - A method used to standardize the range of features of data

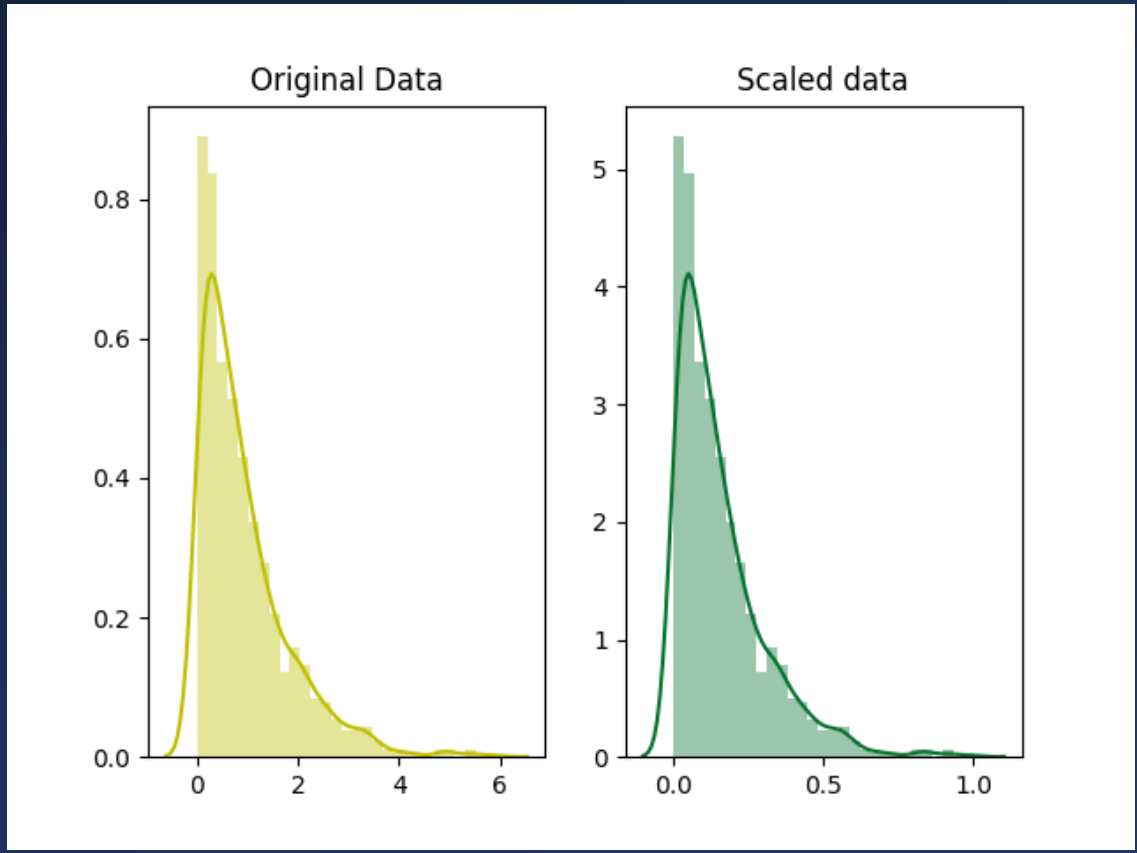
- Scaling features to lie between a given minimum and maximum value

  e.g. between zero and one

  so that the maximum absolute value of each feature is scaled to unit size

- In statistics, scaling usually means a linear transformation

$$f(x) = ax + b$$



- In some machine learning algorithms (e.g., KNN, SVM), objective functions will not work properly without scaling/normalization.
  - e.g. distance based classifiers

$$\sqrt{(x1-x2)^2+(y1-y2)^2}$$

$$\sqrt{(a1-a2)^2+(b1-b2)^2+(c1-c2)^2+(d1-d2)^2+....}$$

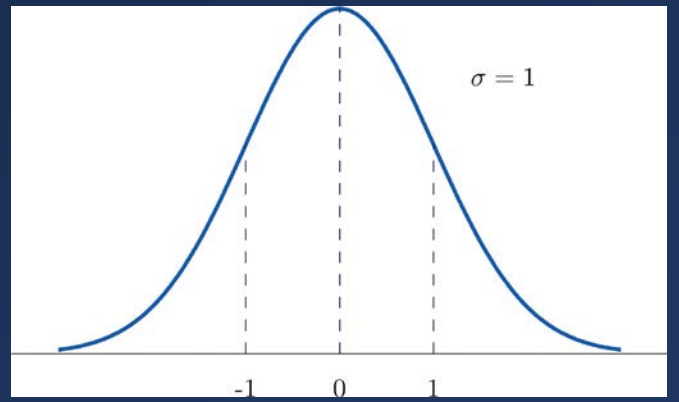
If one of the features has a broad range of values, the distance will be governed by this particular feature.

Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Normalization

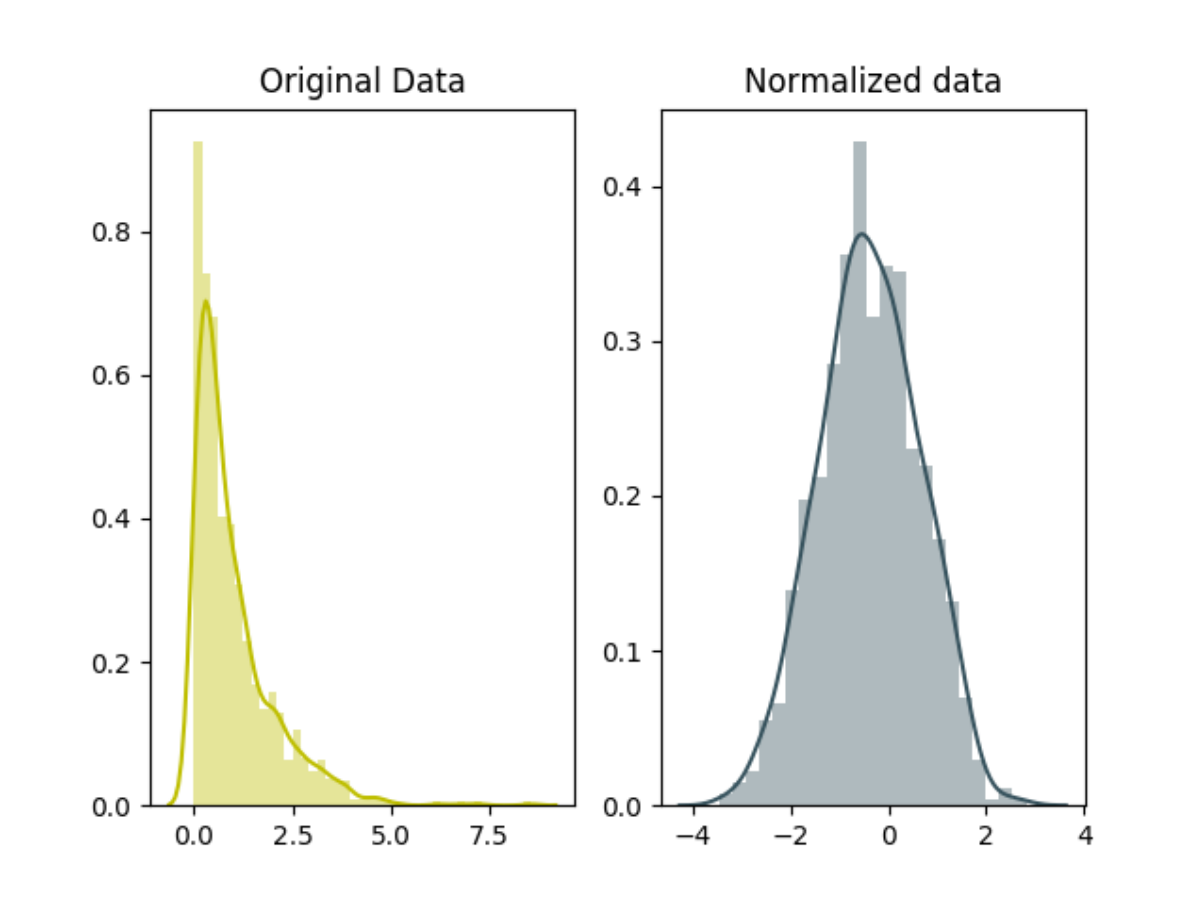
- Z-score normalization, or standardization
- The point of normalization is to change your observations so that they can be described as a normal distribution.
- Transforms your data such that the resulting distribution has the properties of a standard normal distribution with

$$\mu = 0 \text{ and } \sigma = 1$$



- Standard scores (also called ***z*** scores) are calculated by subtracting the mean and dividing by the standard deviation:

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



- Normalization is not only important if we are comparing measurements that have different units, but it is also a general requirement for many machine learning algorithms, e.g., linear regression, Naive Bayes, etc.

- In scaling, you’re changing the range of your data while in normalization you’re mostly changing the shape of the distribution of your data.

- Scaling or Normalization?
  - It really depends on the application
  - Some machine learning or statistics techniques assume that data is normally distributed, e.g., Principal Component Analysis – Normalization
  - Image processing - Scaling pixel intensities

Feature selection

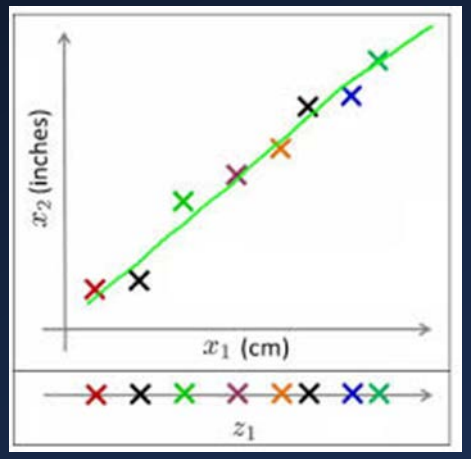
- Also known as variable selection, attribute selection or variable subset selection
- The process of selecting a subset of relevant features (variables, predictors) for use in model construction. In other words, the selection of the most important features
- Top reasons to use feature selection are:
  - It enables the machine learning algorithm to train faster.
  - It reduces the complexity of a model and makes it easier to interpret.
  - It improves the accuracy of a model if the right subset is chosen.
  - It reduces overfitting.

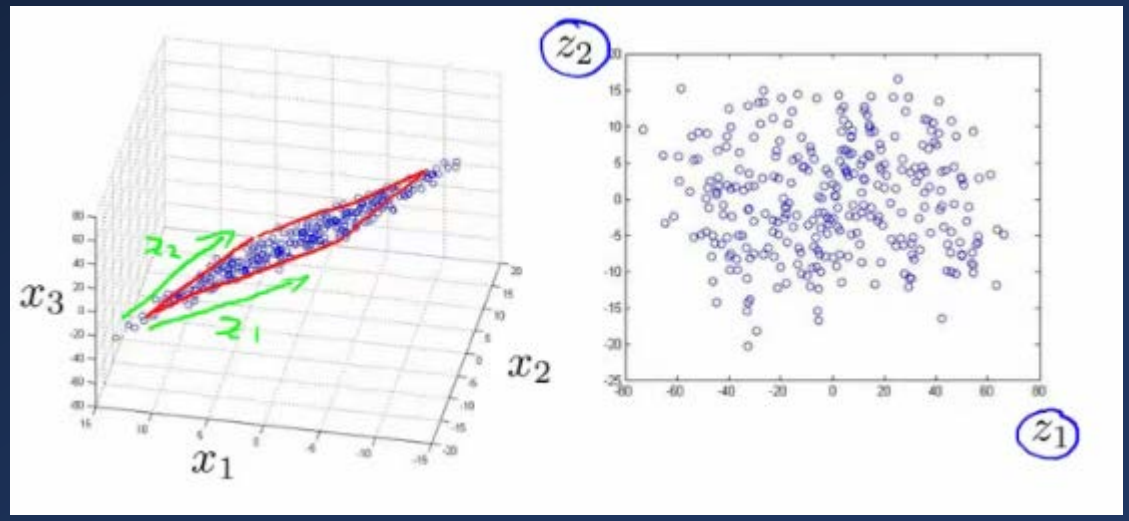
Feature Selection techniques

- Remove features with missing values
  - remove columns/features with missing values exceeding a threshold we define
- Remove features with low variance
  - removes features with zero variance or features that have the same value for all samples
- Remove highly correlated features
  - When a pair of independent variables are highly correlated we can remove one to reduce dimensionality without much loss of information.

Dimensionality reduction

- Also known as dimension reduction, the process of converting a set of data having vast dimensions into data with lesser dimensions ensuring that it conveys similar information concisely.





Principle Component Analysis (PCA)

- PCA is a dimensionality reduction technique that projects the data into a lower dimensional space.

- PCA might be the most popular technique for dimensionality reduction.

- For further details: https://medium.com/analytics-vidhya/dimensionality-reduction-principal-component-analysis-d1402b58feb1

Feature Selection vs Dimensionality Reduction

- Both methods seek to reduce the number of attributes in the dataset.
- Feature selection methods include and exclude attributes present in the data without changing them.
- Dimensionality reduction transforms features into a lower dimension.

Summary

- Missing values
- Invalid data
- Outlier
- Scaling
- Normalization
- Feature selection
- Dimensionality reduction