Deloitte.

Citizen Data Scientist – Part 1 | Class 6
EDA – Exploratory Data Analysis

Instructor: Kamilla Silva

April 2024

MAKING AN IMPACT THAT MATTERS since 1845

Recalling the last class...

Missing values

Main objectives:

- Identifying Missingness Patterns
- Assessing Data Completeness
- **Understanding Missingness Mechanisms**: Different missingness mechanisms, such as missing completely at random (**MCAR**), missing at random (**MAR**), or missing not at random (**MNAR**), require different handling approaches.
- Evaluating Impact on Analysis
- Implementing Handling Strategies

Why look for Missing Values?

- Identify important information that was lost
- Prepare Variable for model

How to solve?

Complete Case Analysis

or

Mean/Median Imputation

or

KNN Imputation

and

Interative Imputation

Outliers

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." [D. Hawkins. Identification of Outliers, Chapman and Hall , 1980.]

Methods that help to identify Outliers:

If the variable is Normally distributed (Gaussian):
• Outliers = mean +/- 3* std

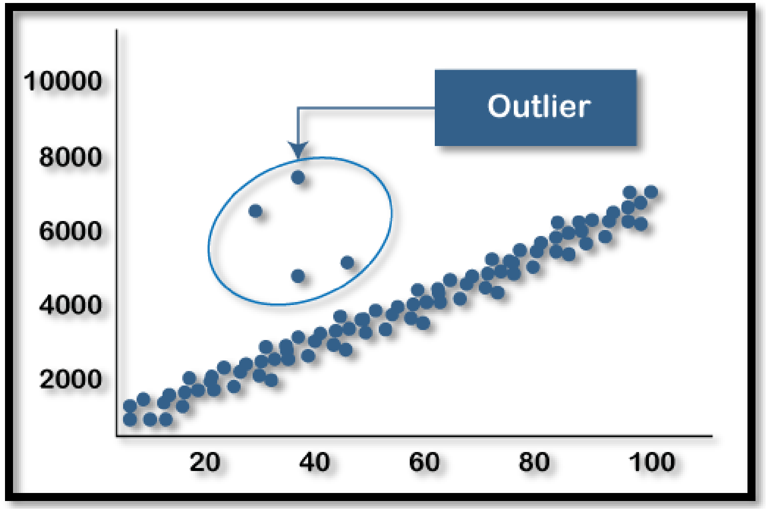
If the variable is skewed distributed, a general approach is to calculate the quantiles, and then the inter-quantile range (IQR), as follows:
•IQR = 75th quantile - 25th quantile

An outlier will sit outside the following upper and lower boundaries:
•Upper boundary = 75th quantile + (IQR * 1.5)
•Lower boundary = 25th quantile - (IQR * 1.5)

or for extreme cases:
•Upper boundary = 75th quantile + (IQR * 3)
•Lower boundary = 25th quantile - (IQR * 3)

Why look for Outiliers?

- Identify suspicious information
- Prepare Variable for model



How to solve?

Trimming
or
Censoring
or
Discretization

* Considering equal width discretization

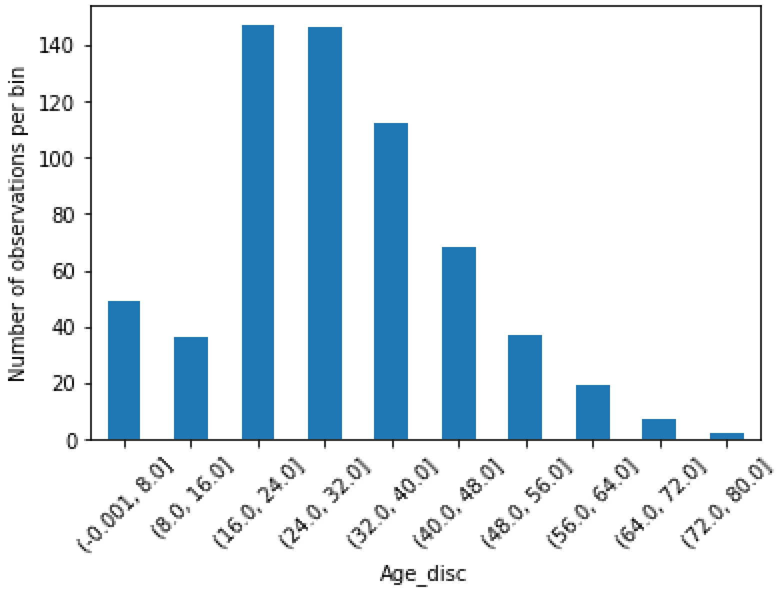
Today’s Class.....

Discretization | Part 2

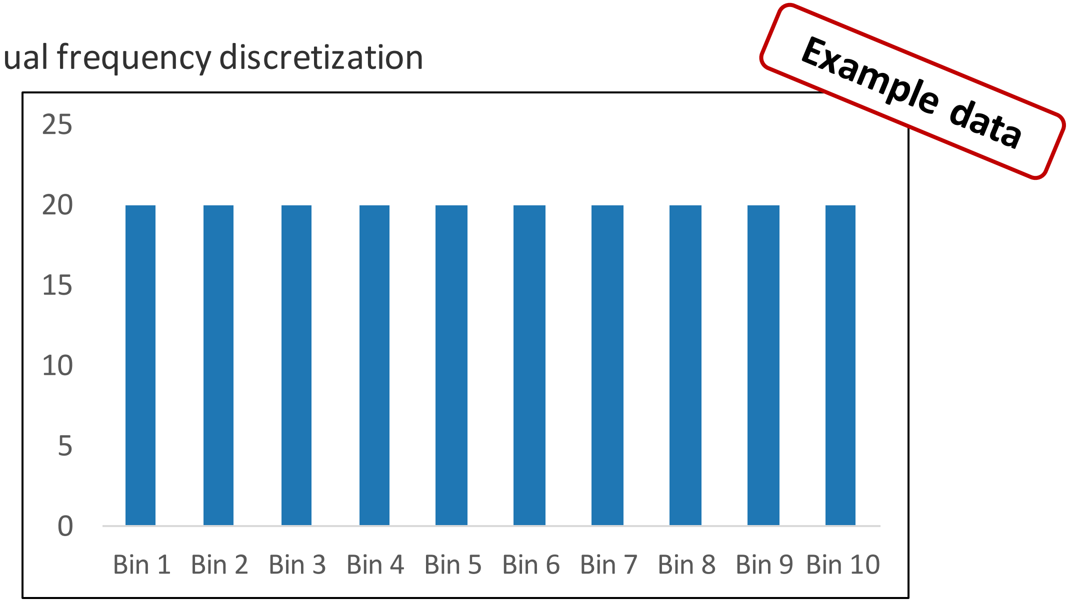
- Considering **equal frequency discretization**;
- Each “Bin” has the same N° of observations;
- Application Method: quantiles.

Example: **Titanic dataset**

- Equal width discretization



- Equal frequency discretization



Feature Engineering

- Standardization
- Normalization
- MinMaxScaling
- Categorical to dummy variables

Feature Scaling

- Why is it importante?

  - The regression coefficients of linear models are directly influenced by the scale of the variable.
  - Variables with bigger magnitude / larger value range dominate over those with smaller magnitude / value range
  - Gradient descent converges faster when features are on similar scales
  - Feature scaling helps decrease the time to find support vectors for SVMs
  - Euclidean distances are sensitive to feature magnitude.
  - Some algorithms, like PCA require the features to be centered at 0.

Feature Engineering

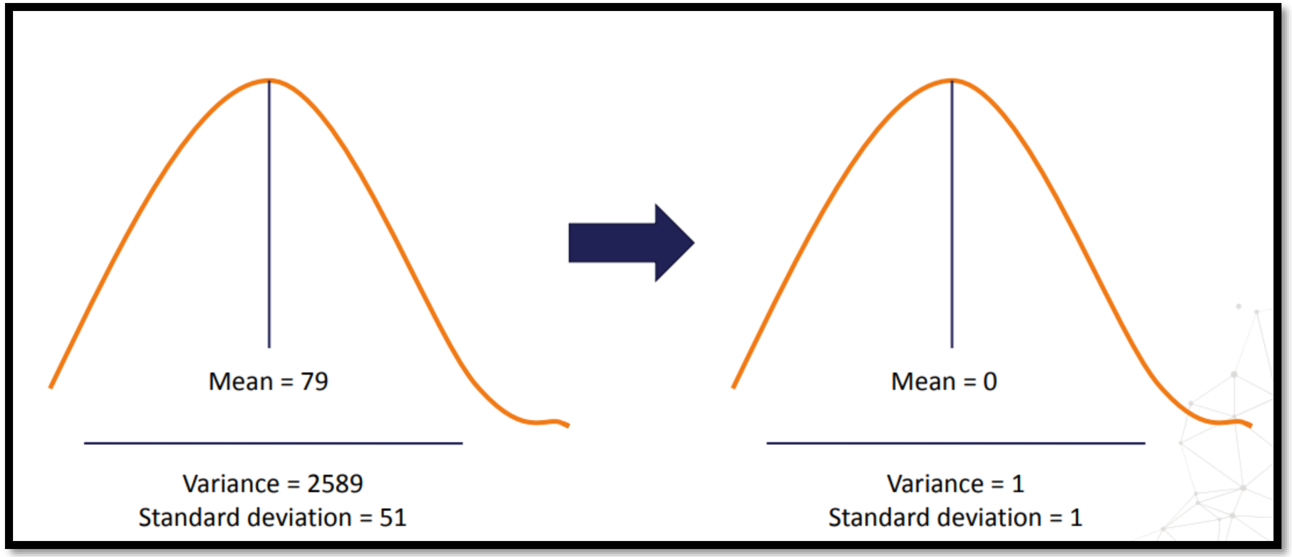
- Standardization
- Normalization
- MinMaxScaling
- Categorical to dummy variables

Feature Scaling

Centres the variable at zero and sets the variance to 1.

$$Z - Score = \frac{x - Mean(X)}{Std(X)}$$

Efect:



Feature Engineering

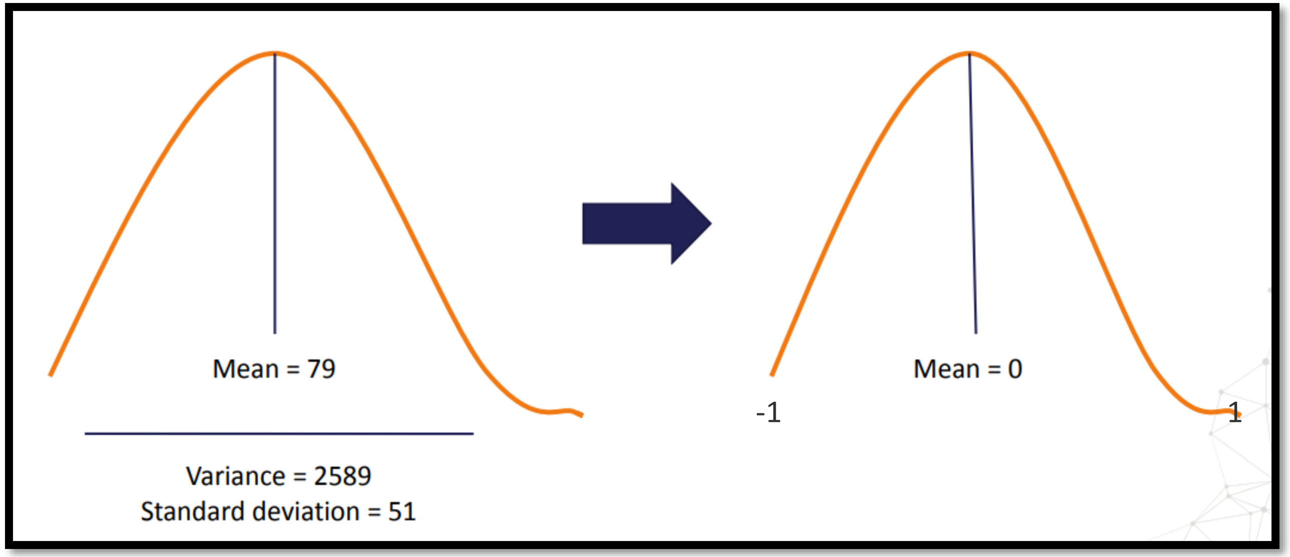
- Standardization
- Normalization
- MinMaxScaling
- Categorical to dummy variables

Feature Scaling

Centres the variable at zero and re-scale the Variable in the value range.

$$X{-}Scaled = \frac{x - Mean\,(X)}{Max\,(X) - Min\,(X)}$$

Efect:



Feature Engineering

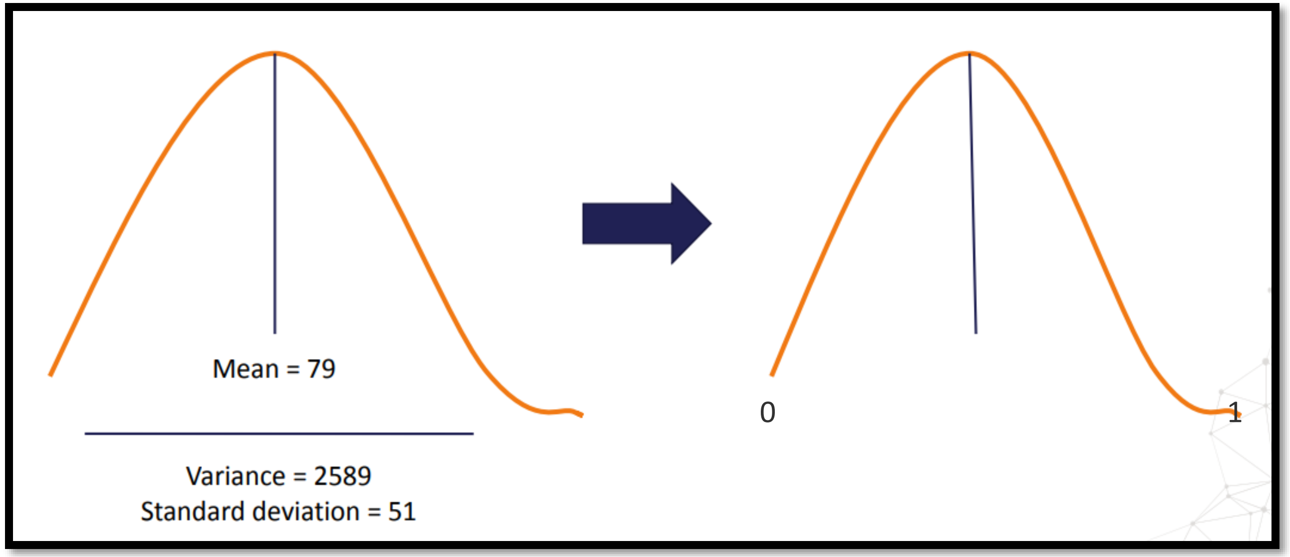
- Standardization
- Normalization
- MinMaxScaling
- Categorical to dummy variables

Feature Scaling

Scales de Variable between 0 and 1.

$$Z - Scaled = \frac{x \ - Min(X)}{Max(X) - \ Min(X)}$$

Efect:



Feature Engineering

- Standardization
- Normalization
- MinMaxScaling

Feature Scaling

- Categorical to dummy variables

Efect:

PANDAS GET DUMMIES CREATES DUMMY VARIABLES FROM CATEGORICAL DATA

sex
male
female
female
male
male
male
male
female
male

`pd.get_dummies()` →

sex_male	sex_female
1	0
0	1
0	1
1	0
1	0
1	0
1	0
0	1
1	0

- Some machine learning algorithms cannot directly work with categorical data;
- Dummy variables are also known as **binary**, because they can assume just two values: 0 or 1.

Let’s Practice in Python…

Asynchronous Topic

- Unsupervised learning : Clustering
  - K-Means
  - Elbow and Silhouette Methods

- Learning Material:
  - **Path**: 2024 CDS Training > Virtual Classroom Training > Asynchronous Topics > 1. Unsupervised learning

Thank you!

Kamilla Silva