Deloitte.

Citizen Data Scientist – Part 1 | Class 5
EDA – Exploratory Data Analysis

Instructor: Kamilla Silva

April 2024

MAKING AN IMPACT THAT MATTERS since 1845

Recalling the last classes...

Deloitte.

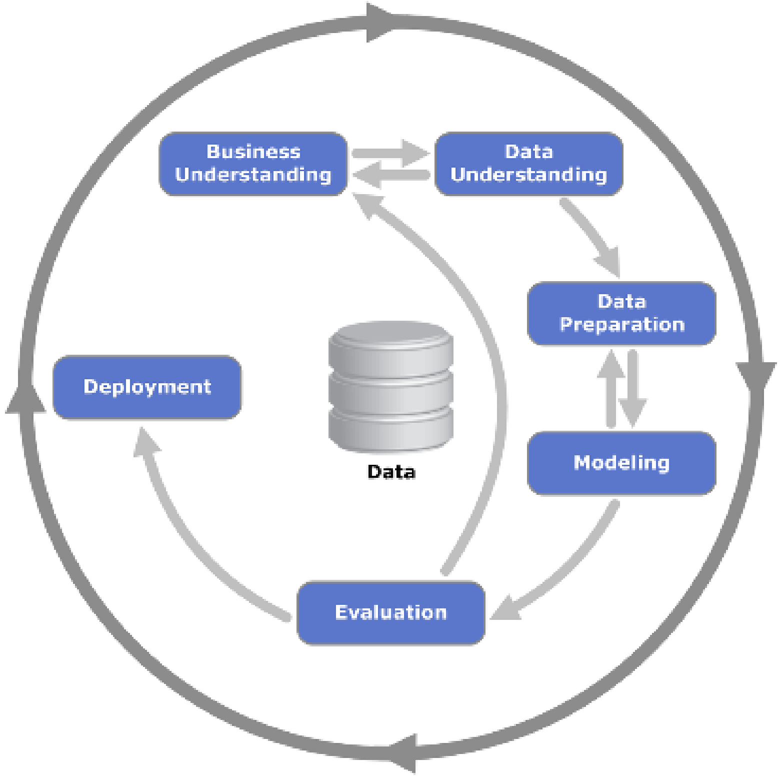
Main Objective of this Course

- Understand and implement techniques for **exploring and analyzing datasets** without predefined labels or targets;
- Gain insights into **data distributions, patterns, and relationships** through visualization and statistical analysis;
- Utilize EDA and unsupervised learning techniques to preprocess and prepare data for further analysis or modeling tasks;
- Interpret and communicate the results of EDA and unsupervised learning analyses effectively to support decision-making processes in various domains.

Exploratory Data Analysis

CRISP-DM

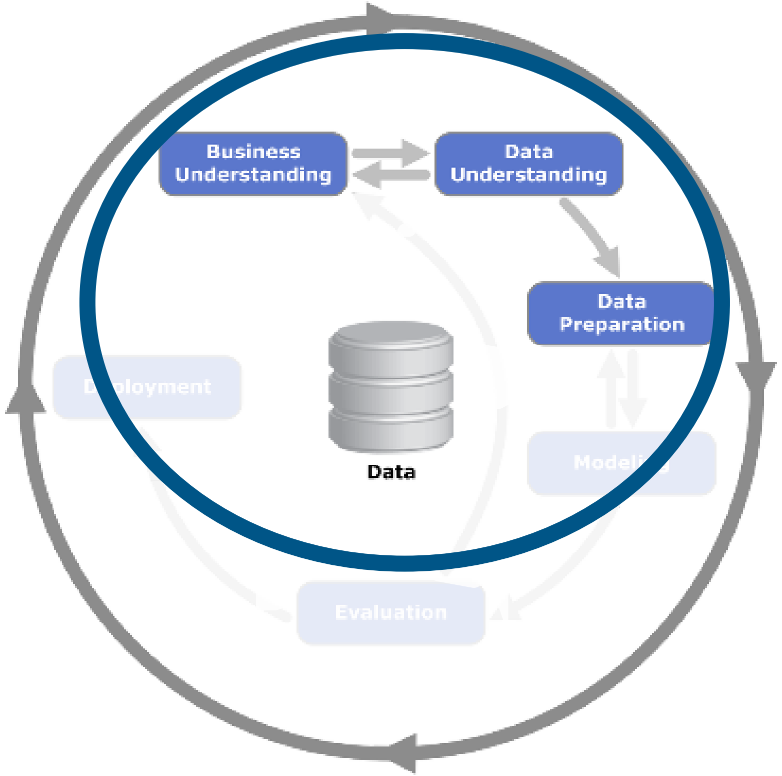
Cross-Industry Standard Process for Data Mining



Structured approach for planning and executing data mining projects.

CRISP-DM

Cross-Industry Standard Process for Data Mining



The purpose of **Exploratory Data Analysis (EDA)** is to summarize the main characteristics of a dataset to better understand its structure, patterns, and relationships.

EDA Steps

Data Collection: Gather the dataset you want to explore.

Data Cleaning: Check for and fix any mistakes or missing values in the data.

Data Exploration: Get a basic understanding of your data through graphs and summary statistics.

Feature Engineering: Create new features or transform existing ones if needed.

Univariate Analysis: Look at individual variables one by one.

Bivariate Analysis: Explore relationships between pairs of variables.

Multivariate Analysis: Examine interactions between multiple variables.

Outlier Detection: Identify and handle any unusual data points that could skew your analysis.

Data Transformation: Prepare the data for modeling by scaling or normalizing it if necessary.

Summary and Insights: Summarize your findings and insights from the EDA process.

Python libraries for data science

NumPy

- Vectors; Matrices
- Documentation

Pandas

- Data frames; Handling tools
- Documentation

Scipy

- Mathematical calculations and statistics
- Documentation

Matplotlib

- Charts; Images
- Documentation

Seaborn

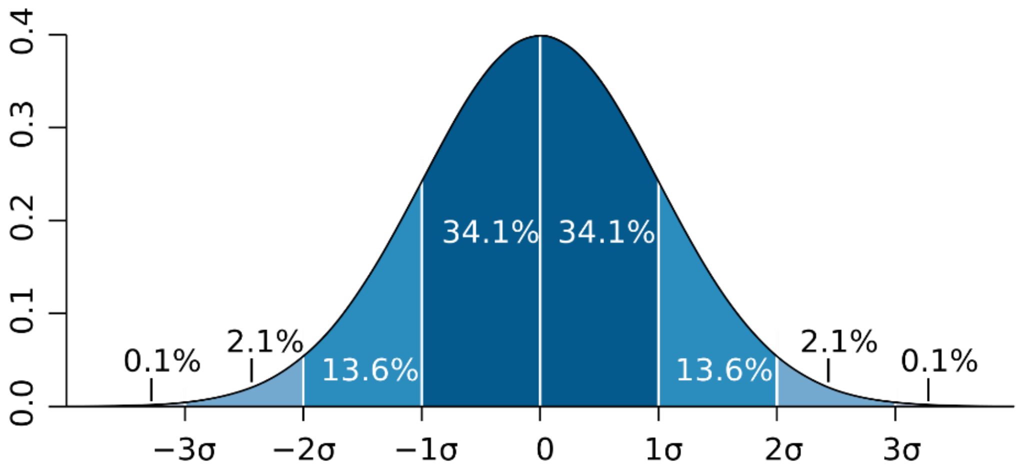
- Enhanced charts; Exploratory Data Analysis
- Documentation

Scikit-learn

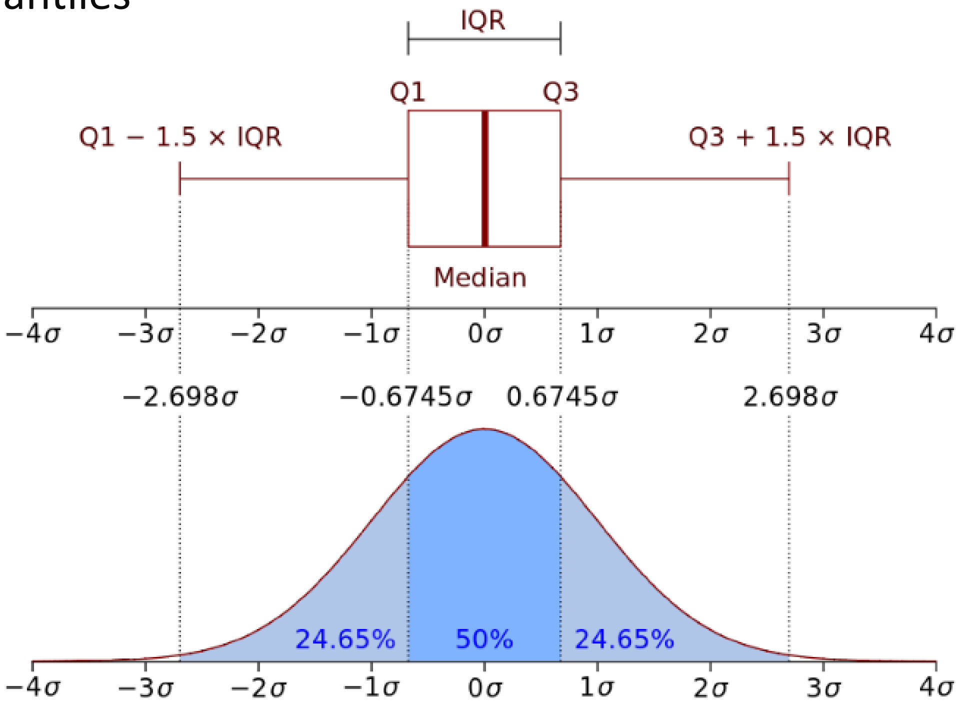
- Machine learning
- Documentation

Statistical Analysis

- Normal Distribution



- Quantiles



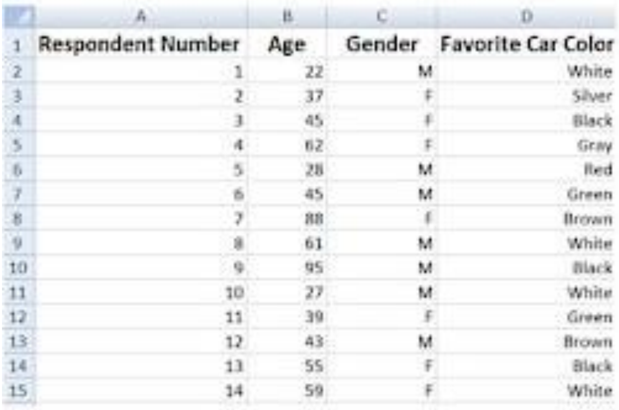
EDA - Hands On

Descriptive Analysis

with Pandas

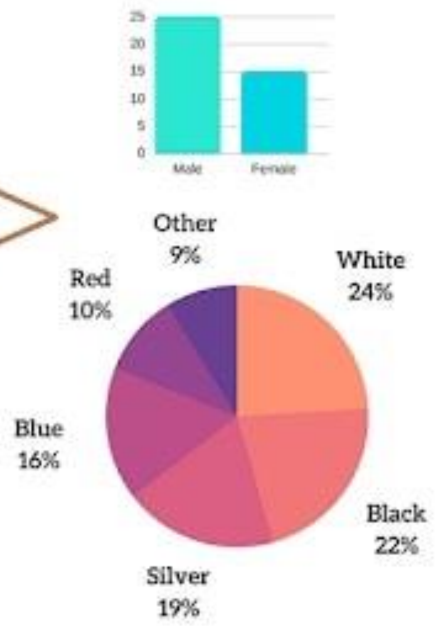
Main objectives:

- Understanding data
- Data Summarization
- Data Exploration
- Quality Assessment



	A	B	C	D
1	Respondent Number	Age	Gender	Favorite Car Color
2	1	22	M	White
3	2	37	F	Silver
4	3	45	F	Black
5	4	62	F	Gray
6	5	28	M	Red
7	6	45	M	Green
8	7	88	F	Brown
9	8	61	M	White
10	9	95	M	Black
11	10	27	M	White
12	11	39	F	Green
13	12	43	M	Brown
14	13	55	F	Black
15	14	59	F	White

RAW DATA



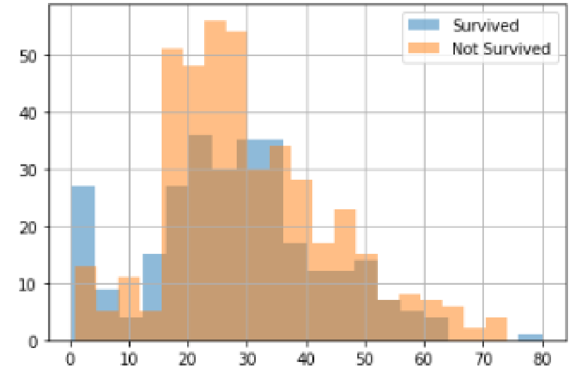
Descriptive Statistics

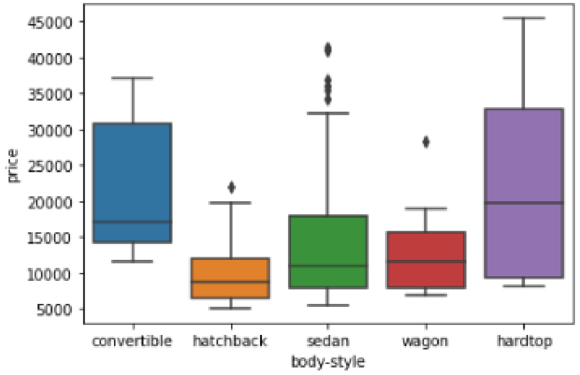
Plot Visualization

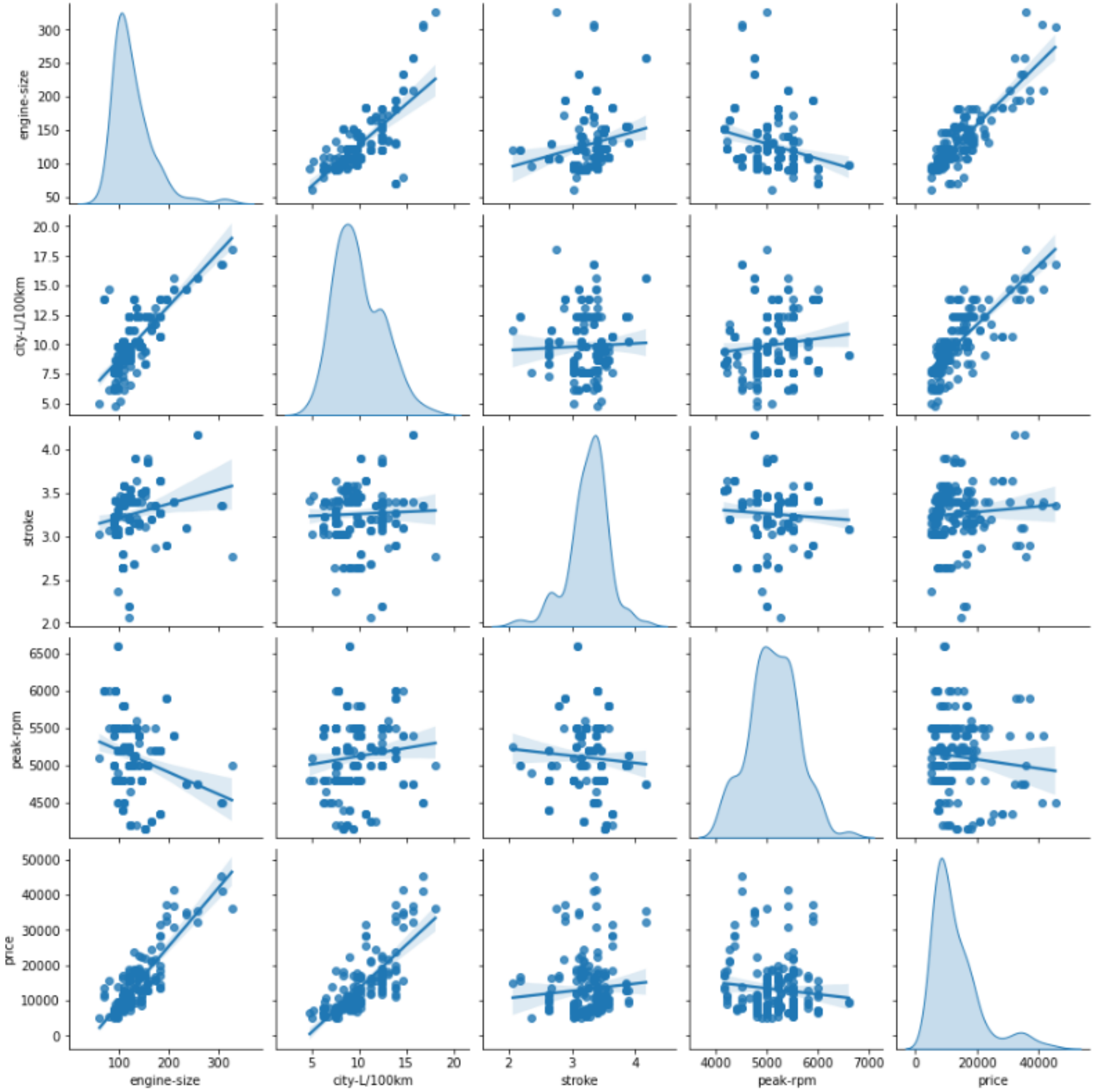
with Seaborn

Main objectives:

- Understanding data
- Indentifying Relationships
- Spotting trends and patterns
- Comparing groups or categories
- Comunicating Insights

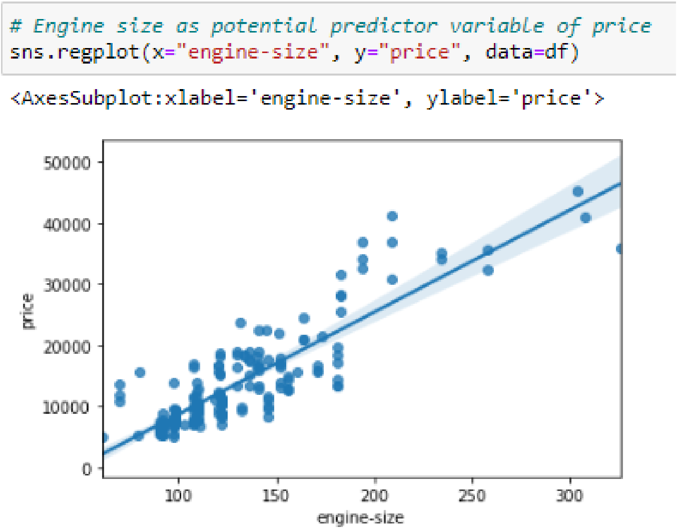






Correlations

with Stats using Pearson’s Correlation and ANOVA



```
# Engine size as potential predictor variable of price
sns.regplot(x="engine-size", y="price", data=df)
```

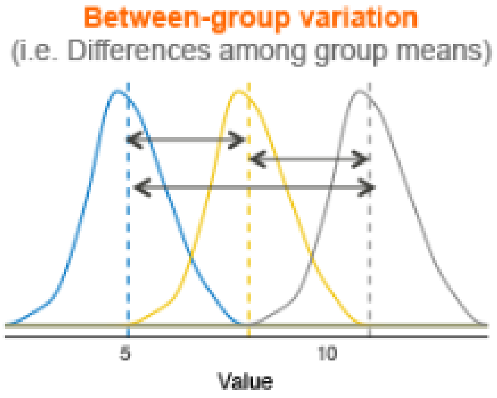
```
<AxesSubplot:xlabel='engine-size', ylabel='price'>
```

Pearson's Correlation Analysis:

- Pearson's correlation analysis is used to measure the strength and direction of the linear relationship between two continuous variables.
- Its main objective is to assess the degree of association between variables, indicating how changes in one variable are related to changes in another variable.
- Pearson's correlation coefficient (r) ranges from -1 to +1, where values close to +1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values close to 0 indicate no linear correlation.

ANOVA (Analysis of Variance):

- ANOVA is primarily used to analyze the differences among means of three or more groups or treatments.
- Its main objective is to determine whether there are statistically significant differences between the means of the groups being compared.
- ANOVA helps in understanding the impact of categorical independent variables on a continuous dependent variable and identifying which groups differ significantly from each other.



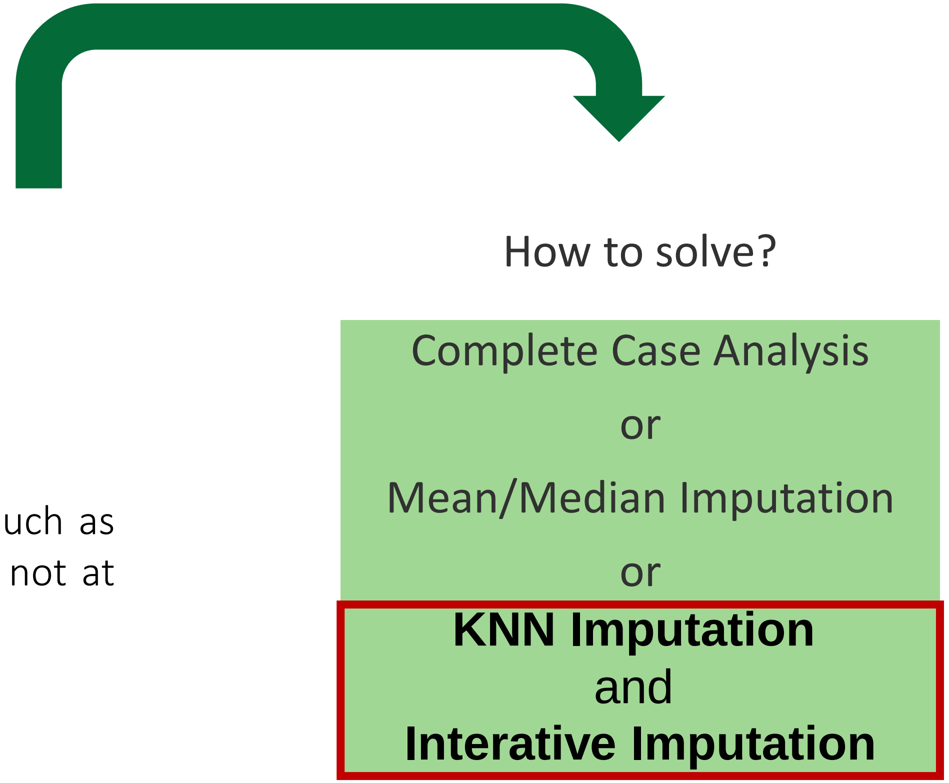
Why look for Correlations?

- Indentify crucial predictors
- Parsimony

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



How to solve?

Complete Case Analysis

or

Mean/Median Imputation

or

KNN Imputation
and
Interative Imputation

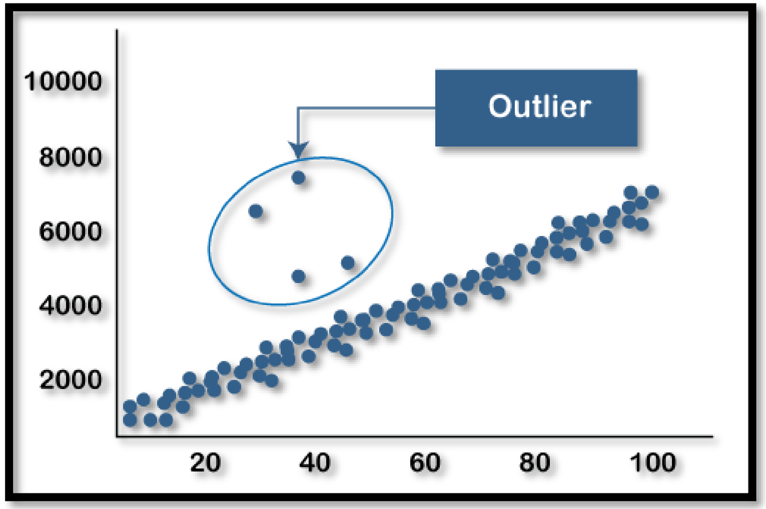
Today’s class

Why look for Missing Values?

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

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

Today's class

Recommended Reading

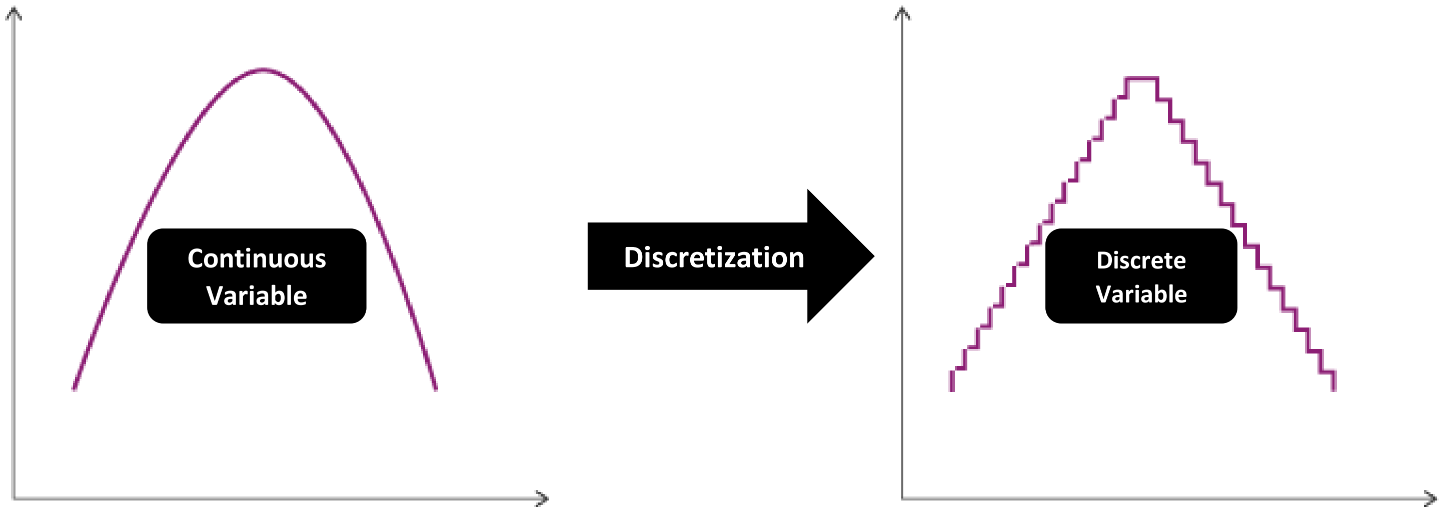
Machine Learning University:

- https://mlu-explain.github.io/

Today’s Class…..

Discretization

- The process of transforming continuous functions, models, variables and equations into discrete counterparts.
- This is important, as some algorithms only work with inputs of discrete values, not predicting continuous values.
- Discretization creates a limited number of possible states.



Discretization

Example: Car Insurance Price

- Age is a good predictor for the risk of na accident
- There is no significant difference in risk for individuals aged 18 or 19
- Creating age groups helps in separating the risk

Probability of an accident occurring by groups:

- Group 1: Individuals aged 18-25 years
- Group 2: Individuals aged 26-30 years
- Group 3: Individuals aged 30-45 years
- Group 4: Individuals aged 45-60 years
- Group 5: Individuals aged over 60 years