DATA SCIENCE BOOTCAMP

Session 8: EDA – Data Preparation

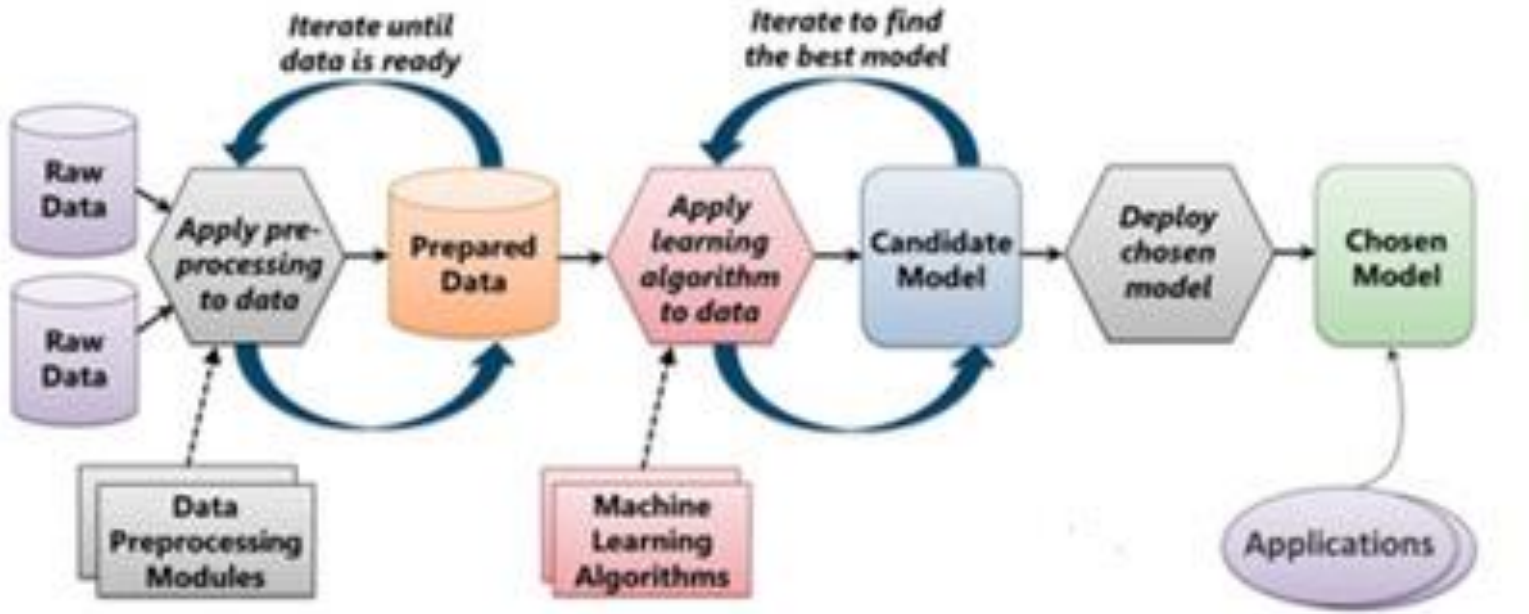
SESSION 8 OUTLINE

- MACHINE LEARNING PROCESS RECAP
- DESCRIPTIVE STATISTICS
- VERIFYING DATA QUALITY: COMMON ISSUES
- CONVERTING: CATEGORY TO NUMBERS
- SCALING

SESSION 8 OUTLINE

- MACHINE LEARNING PROCESS RECAP
- DESCRIPTIVE STATISTICS
- VERIFYING DATA QUALITY: COMMON ISSUES
- CONVERTING: CATEGORY TO NUMBERS
- SCALING

MACHINE LEARNING PROCESS



From "Introduction to Microsoft Azure" by David Chappell

Google Images: "machine learning process"

SESSION 8 OUTLINE

- MACHINE LEARNING PROCESS RECAP
- DESCRIPTIVE STATISTICS
- VERIFYING DATA QUALITY: COMMON ISSUES
- CONVERTING: CATEGORY TO NUMBERS
- SCALING

DESCRIPTIVE STATISTICS

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread)

https://www.investopedia.com/

DESCRIPTIVE STATISTICS

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread)

https://www.investopedia.com/

DESCRIPTIVE STATISTICS

.head()
.tail()
.shape
.columns
.info()
.dtypes
.describe()

SESSION 8 OUTLINE

- MACHINE LEARNING PROCESS RECAP
- DESCRIPTIVE STATISTICS
- VERIFYING DATA QUALITY: COMMON ISSUES
- CONVERTING: CATEGORY TO NUMBERS
- SCALING

DATA QUALITY COMMON ISSUES

- Improper Labels
- Missing Values
- Incorrect Formats
- Duplicated Data
- Incorrect Spelling
- Outliers

DATA QUALITY COMMON ISSUES

- Improper Labels
- Missing Values
- Incorrect Formats
- Duplicated Data
- Incorrect Spelling
- Outliers

Id	0	Br4nd
STP-489762	Chocolate	Almond
STP-489763	Toothpaste	Colgate
STP-489764	Shampoo	Hair and Shoulders

O -> Product
Br4nd -> Brand

DATA QUALITY COMMON ISSUES

- Improper Labels
- Missing Values
- Incorrect Formats
- Duplicated Data
- Incorrect Spelling
- Outliers

Id	0	Br4nd
STP-489762	Chocolate	Almond
STP-489763		Colgate
STP-489764		Hair and Shoulders

** drop the column or handle using other values – mean, mode*

DATA QUALITY COMMON ISSUES

- Improper Labels
- Missing Values
- Incorrect Formats
- Duplicated Data
- Incorrect Spelling
- Outliers

```
0   Date                16834 non-null  object
1   product             16834 non-null  object
2   phase               16834 non-null  object
3   campaign_platform   16834 non-null  object
4   campaign_type       16834 non-null  object
```

Convert to datetime instead of object
Convert to numbers instead of object

DATA QUALITY COMMON ISSUES

- Improper Labels
- Missing Values
- Incorrect Formats
- **Duplicated Data**
- Incorrect Spelling
- Outliers

Id	0	Br4nd
STP-489762	Chocolate	Almond
STP-489762	Chocolate	Almond
STP-489764	Shampoo	Hair and Shoulders

.duplicate()

DATA QUALITY COMMON ISSUES

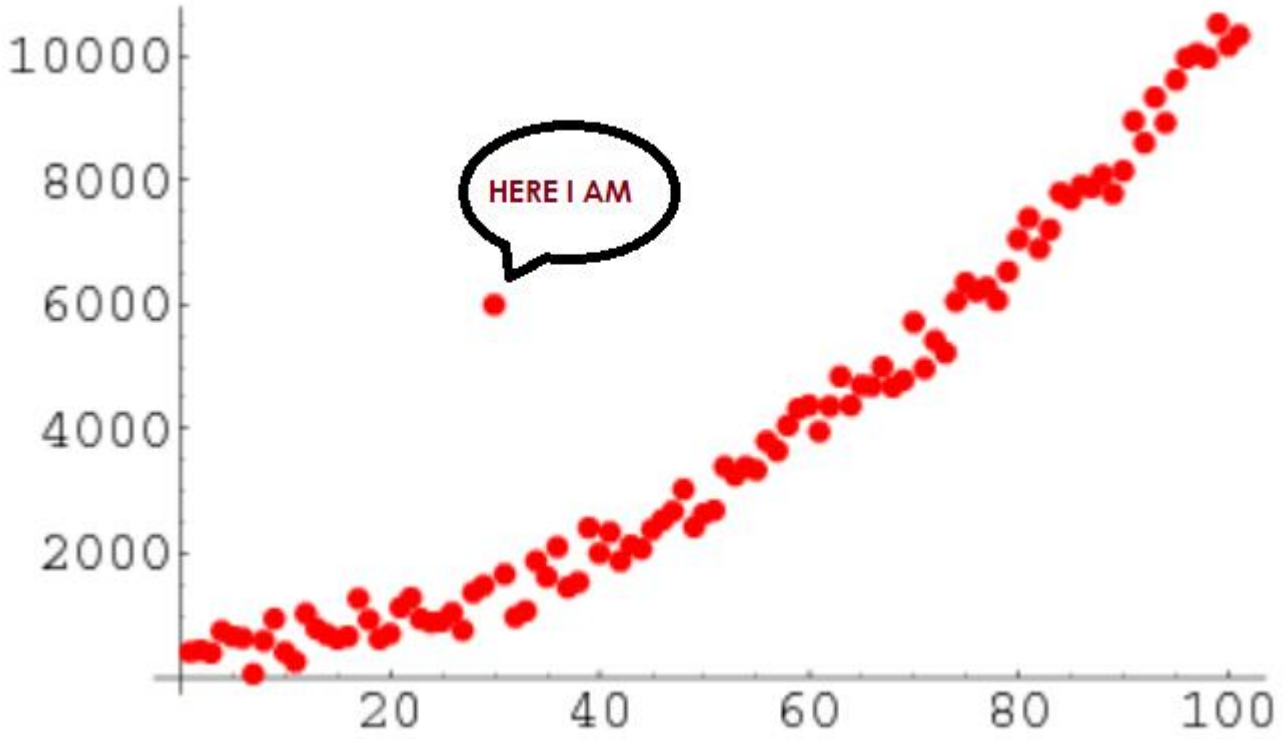
- Improper Labels
- Missing Values
- Incorrect Formats
- Duplicated Data
- **Incorrect Spelling**
- Outliers

Id	0	Br4nd
STP-489762	Chokoleyt	Almond
STP-489762	Chocolate	Almond
STP-489764	Shampoo	Hair and Shoulders

** Replace the values with correct spelling*

DATA QUALITY COMMON ISSUES

- Improper Labels
- Missing Values
- Incorrect Formats
- Duplicated Data
- Incorrect Spelling
- **Outliers**



Google Images: "outliers"

SESSION 8 OUTLINE

- MACHINE LEARNING PROCESS RECAP
- DESCRIPTIVE STATISTICS
- VERIFYING DATA QUALITY: COMMON ISSUES
- CONVERTING: CATEGORY TO NUMBERS
- SCALING

CONVERTING: CATEGORY TO NUMBERS

Pikachu	→	1
Bulbasaur	→	2
Squirtle	→	3

Remember, a computer can only understand 1s and 0s

Google Images: "computer 1 and 0"

SESSION 8 OUTLINE

- MACHINE LEARNING PROCESS RECAP
- DESCRIPTIVE STATISTICS
- VERIFYING DATA QUALITY: COMMON ISSUES
- CONVERTING: CATEGORY TO NUMBERS
- SCALING

SCALING

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.

It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

https://www.geeksforgeeks.org/

SCALING

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.

It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

https://www.geeksforgeeks.org/

DATA SCIENCE BOOTCAMP

Session 8: EDA – Data Preparation