LO5

Prepare Data for Modeling

Objective

After attending this session, you should know

- Treating missing values
- Removing duplicates
- Concatenating and transforming
- Grouping and aggregation

Missing value in python

- By default, missing values are represented in Python with NaN

Respondent Name	Speed of Service (1–10)	Friendliness of Staff (1–10)	Quality of Work (1–10)	Price (1–10)
Sally	10	10	missing	8
Jim	10	9	missing	10
Rod	9	8	9	9
Sam	8	10	8	10
Jane	10	4	10	6

Use mean to approximate the missing value

Respondent Name	Speed of Service (1–10)	Friendliness of Staff (1–10)	Quality of Work (1–10)	Price (1–10)
Sally	10	10	**8**	8
Jim	10	9	**8**	10
Rod	9	8	9	9
Sam	8	10	8	10
Jane	10	4	10	6

Missing values in python

- How to discover what's missing
- How to fill in for missing values
- How to count missing values
- How to filter out using missing values

Why remove duplicate duplicate?

- It's really important to remove duplicates from your dataset in order to
  - Preserve the dataset's accuracy
  - Avoid producing incorrect and misleading statistics.

Name	Zip	Credit Card Number
Sally	32803	123456789123
Sally	32803	234567891234
Sally	32803	345678912345

Concatenation & Data transformation

- Useful for getting your data into the structure and order you need for analysis.
- Concatenating is simply combining data from separate sources
- Transformation is converting and reformatting data to the format that's necessary for your purposes

Subgrouping your data

- Grouping and aggregation are useful for exploring and describing your dataset in its subgroups.
- Grouping is an excellent method to use when you want to explore and understand your data and its inherent subgroups.

Index	Fruit
1	Apple
2	Apple
3	Orange
4	Apple
5	Orange

Sub-Group
Apple
Orange

- You can group data, in order to
  - compare subsets
  - deduce reasons why subgroups differ the way they do
  - may only be interested in specific subgroups for your analysis. Grouping can help you identify and subset out those subgroups

Summary

- Treating missing values
- Removing duplicates
- Concatenating and transforming
- Grouping and aggregation

Himanshu Patel, Instructor
Saskatchewan Polytechnic
email: patelh@saskpolytech.ca
Mining building, Saskatoon