IBM Course Seven

Week Two

Objectives

01

Describe data formatting techniques

02

Demonstrate the use of binning and of categorical variables

03

Identify data preprocessing techniques

Describe data normalization

How much do you remember?

Test your knowledge of definitions!

Data Formatting

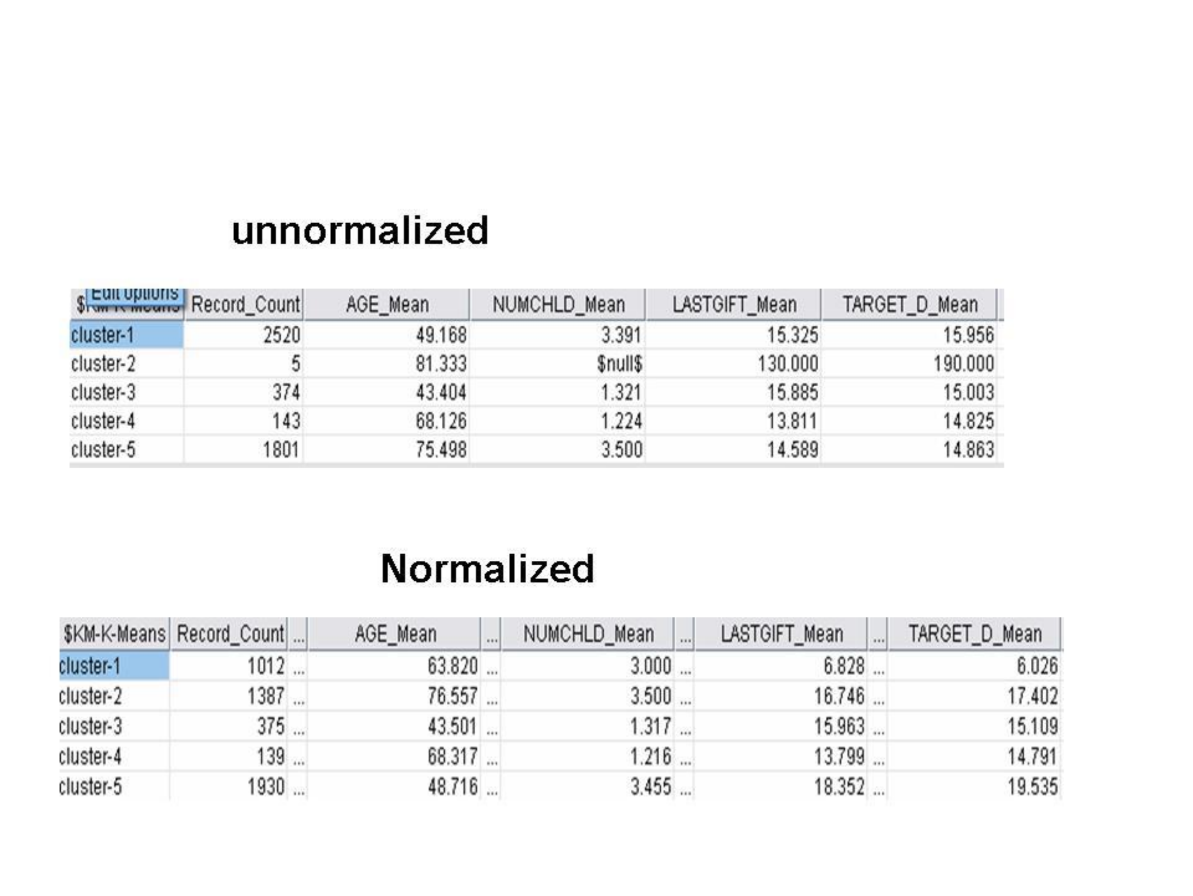
Data are usually collected from different places and stored in different formats.

Bringing data into a common standard of expression allows users to make meaningful comparisons.

Data Types in Python

- Sometimes the wrong data type is assigned to a feature.
- Objects □ letters or words
- Int64 □ integers
- Float64 □ real numbers

    - *What's the difference between an integer and a real number?*



unnormalized

$[illegible]	Record_Count	AGE_Mean	NUMCHLD_Mean	LASTGIFT_Mean	TARGET_D_Mean
cluster-1	2520	49.168	3.391	15.325	15.956
cluster-2	5	81.333	$null$	130.000	190.000
cluster-3	374	43.404	1.321	15.885	15.003
cluster-4	143	68.126	1.224	13.811	14.825
cluster-5	1801	75.498	3.500	14.589	14.863

Normalized

$KM-K-Means	Record_Count	...	AGE_Mean	...	NUMCHLD_Mean	...	LASTGIFT_Mean	...	TARGET_D_Mean
cluster-1	1012	...	63.820	...	3.000	...	6.828	...	6.026
cluster-2	1387	...	76.557	...	3.500	...	16.746	...	17.402
cluster-3	375	...	43.501	...	1.317	...	15.963	...	15.109
cluster-4	139	...	68.317	...	1.216	...	13.799	...	14.791
cluster-5	1930	...	48.716	...	3.455	...	18.352	...	19.535

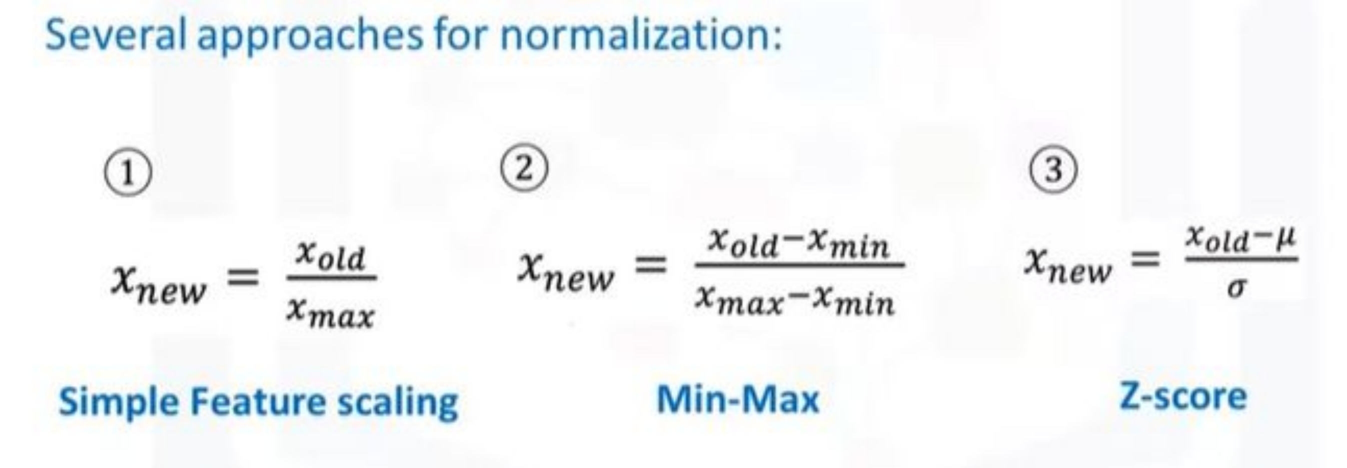
Why is data normalization important?

Ways to Normalize Data

Simple Feature Scaling

Min-Max

Z-score (or Standard Score)



Several approaches for normalization:

①

$$x_{new} = \frac{x_{old}}{x_{max}}$$

Simple Feature scaling

②

$$x_{new} = \frac{x_{old}-x_{min}}{x_{max}-x_{min}}$$

Min-Max

③

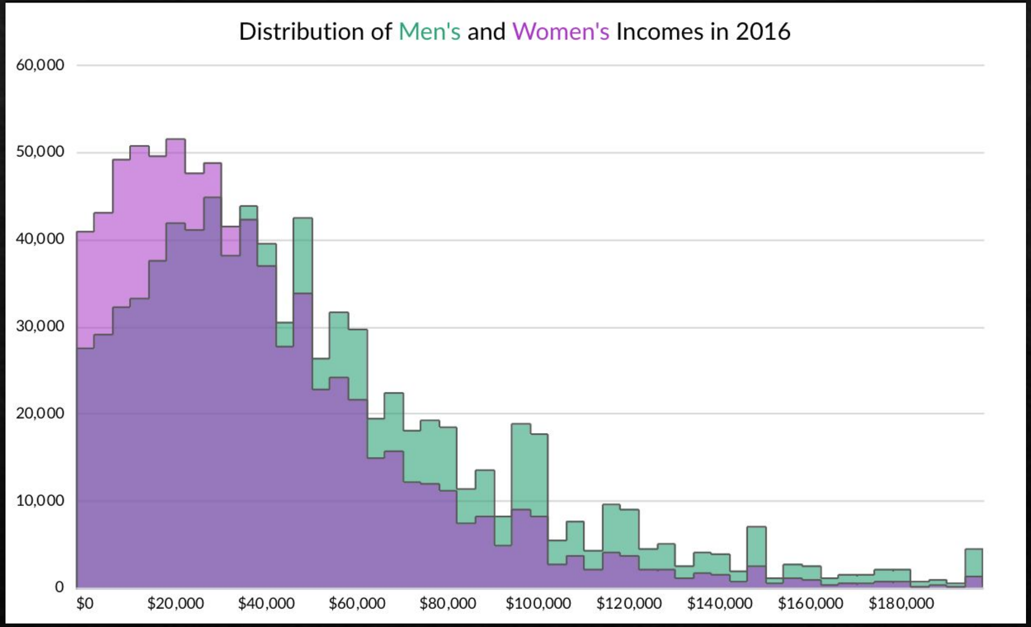
$$x_{new} = \frac{x_{old}-\mu}{\sigma}$$

Z-score

Binning

- Grouping values into bins
- Convert numeric into categorical variables

- *After data has been put into bins, what would be the best graph to use to visualize the data?*



Resources to Explore!

- Real Python
- Repl

Lab and Assessment

- *Please spend the next 2 hours working through the lab and assessment for week 2.*