DATA SCIENCE

TRAINING PROGRAM

Data Manipulation

Larissa Leite
Kodit.io

Outline

- When and why data is manipulated?
- Data cleaning
- Numerical data transformations
  - Feature scaling
  - Logarithmic and exponential
  - Discretization
- Categorical data transformations
  - Label and One-hot encoding
- Textual data transformations
- Feature selection
- Dimensionality reduction

When and why manipulate the data?

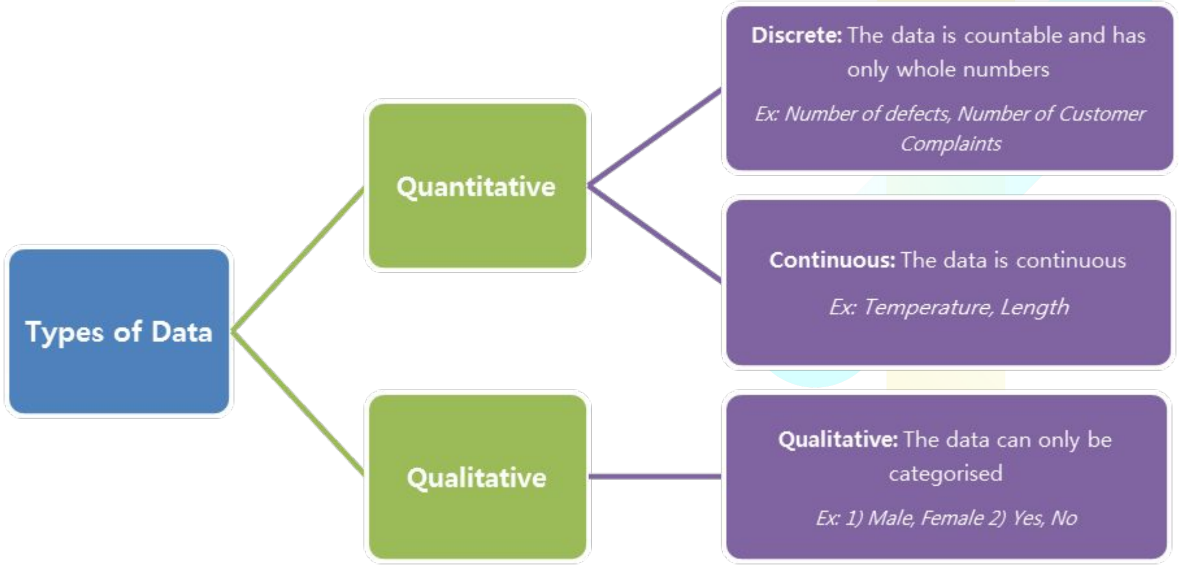
- When exploring the data
  - To understand it
  - Examples:
    - Fill-in missing values
    - Eliminate duplicates and wrong values
    - Change data types or data distribution
    - Reduce the number of dimensions

<u>When</u> and <u>why</u> manipulate the data?

- When preparing the data for a machine learning model
  - To make it compliant with the model's assumptions
    - Normality, linearity, homogeneity of variance
  - To make units of attributes comparable when measured on different scales
    - Elevation ranging from 100 to 2000 meters and slope from 0 to 30 degrees
  - To reduce the effect of total quantity (sample unit totals), focusing on relative quantities
  - To equalize (or otherwise alter) the relative importance of common and rare values
  - To improve a model's performance
- Most datasets will benefit of one or more data transformations

Data types - recap.

- Numerical
  - Continuous
  - Discrete
- Categorical
  - Nominal
  - Ordinal
- Text
- Media



http://www.datasavvies.com/wp-content/uploads/2018/07/Basic-Statistics-Types-of-Data.png

Data cleaning

- **Big part of data science work**

Data cleaning

- **Big part of data science work**

- Outliers detection & removal
  - Monthly income with negative values

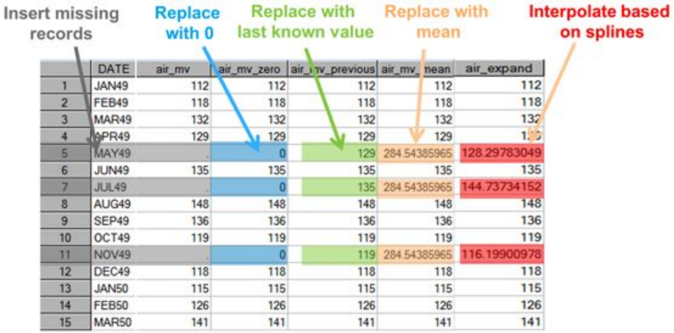
- Especially necessary if the data fields were entered by humans
  - Typos, incorrect or invalid data

the shortcut × mozilla

Data cleaning

- Fill-in missing values
  - How to choose what's appropriate?
    - It really depends on the problem!

  - <u>Numerical</u>: mean, median, 0, -1?
  - <u>Categorical</u>: unknown? Default field? Most common?
    - Develop a predictor to guess?
    - Postcode missing? What about the address?



	DATE	air_mv	air_mv_zero	air_mv_previous	air_mv_mean	air_expand
1	JAN49	112	112	112	112	112
2	FEB49	118	118	118	118	118
3	MAR49	132	132	132	132	132
4	APR49	129	129	129	129	[illegible]
5	MAY49	.	0	129	284.54385965	128.29783049
6	JUN49	135	135	135	135	135
7	JUL49	.	0	135	284.54385965	144.73734152
8	AUG49	148	148	148	148	148
9	SEP49	136	136	136	136	136
10	OCT49	119	119	119	119	119
11	NOV49	.	0	119	284.54385965	116.19900978
12	DEC49	118	118	118	118	118
13	JAN50	115	115	115	115	115
14	FEB50	126	126	126	126	126
15	MAR50	141	141	141	141	141

Data transformations

Numerical data

- Feature scaling (normalization)
  - Min-max (scaling)
  - Z-score (standardization)
- Logarithmic and exponential transformations
- Discretization

Feature scaling

Numerical data

- Min-Max (scaling):
  - Rescale continuous variables to a scale from 0 to 1

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Z-score (standardization):
  - Transform the data to have a mean of 0 and a standard deviation of 1

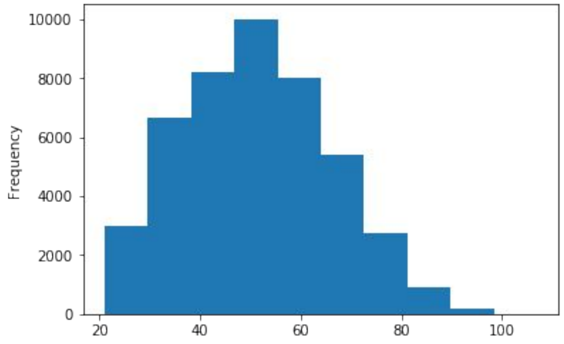
$$z = \frac{x_i - \mu}{\sigma}$$

- When/why?
  - Some machine learning models are affected by the order of magnitude and/or the variance of the input variables, both in terms of results and convergence speed
    - SVM, Neural networks, KNN, K-Means, Logistic Regression
    - Distance-based calculations

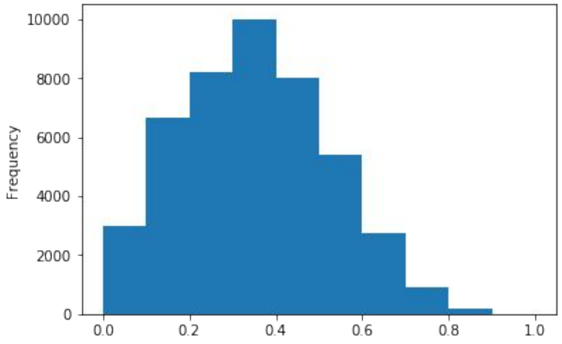
Feature scaling

Numerical data

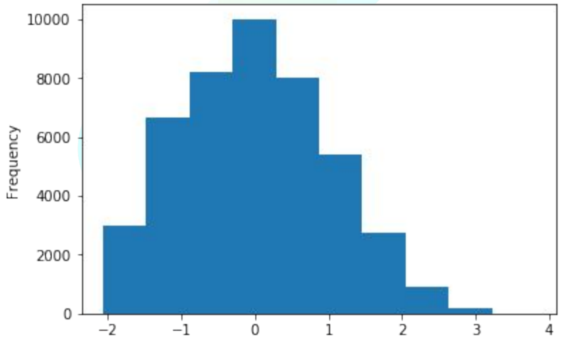
- Age variable



No transformation



Scaling

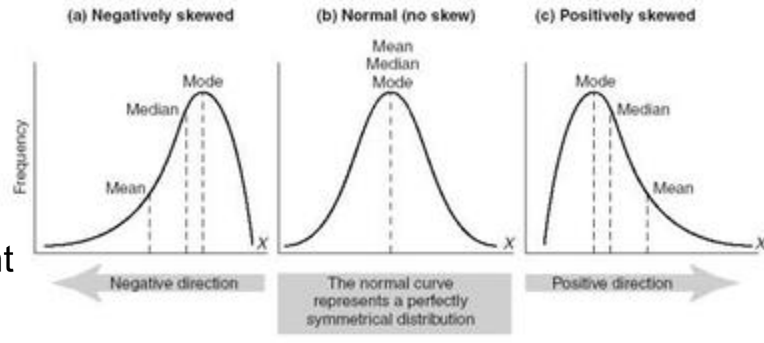


Standardization

Logarithmic and Exponential transformations

Numerical data

- When/why?
  - Skewed data
- Logarithmic
  - Strong transformation that can be used to reduce right skewness
- Square root
  - Medium effect transformation to reduce left skewness
  - Applied to positive values only
- Cube root
  - Fairly strong transformation with a substantial effect on distribution shape
  - It can be applied to negative and zero values

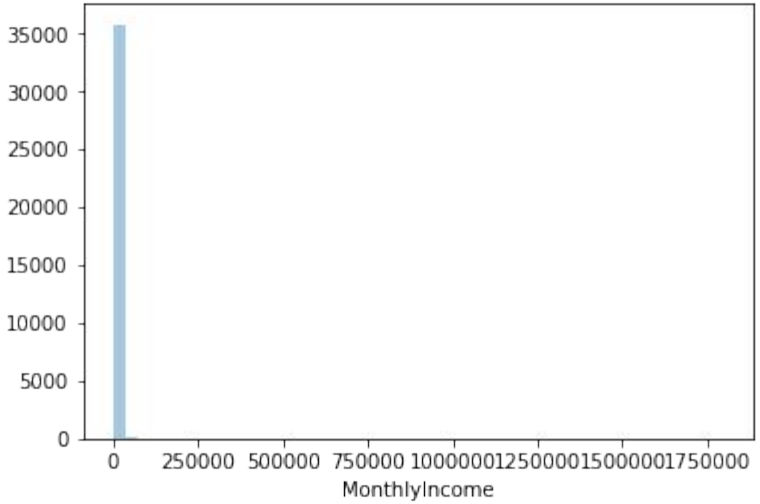


Examples of normal and skewed distributions

Logarithmic and Exponential transformations

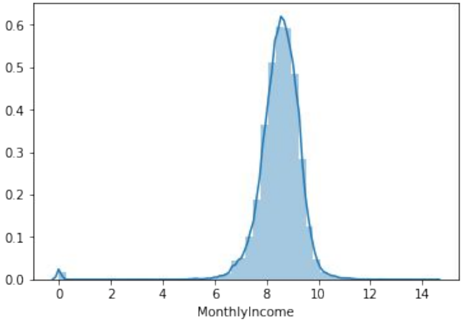
Numerical data

Original distribution

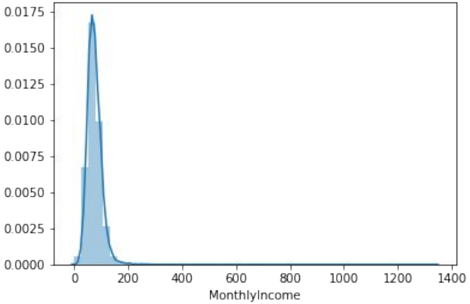


Logarithmic and Exponential transformations

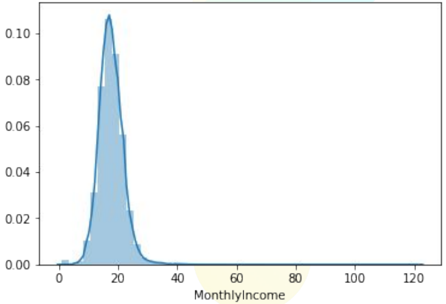
Numerical data



Logarithmic



Square root



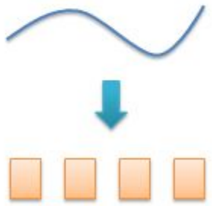
Cube root

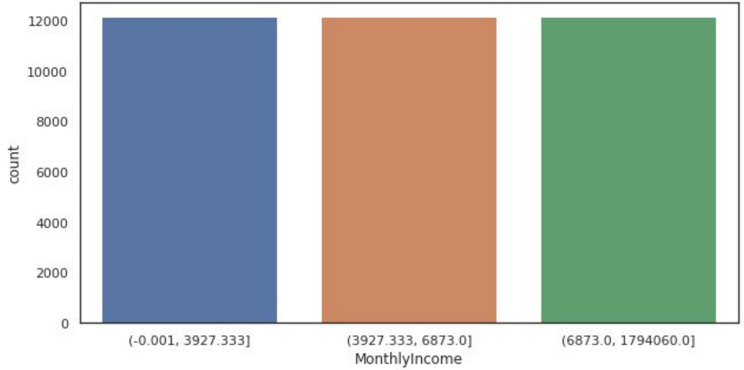
the shortcut × moz://a

Discretization (bucketing)

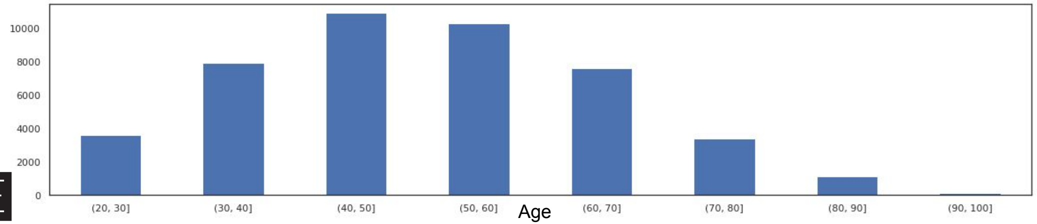
Numerical data

- Discretization consists of 2 steps:
  - Dividing a continuous variable into segments
  - Grouping the segments into bins/categories





- Techniques
  - Equal width (equally spaced boundaries)
  - Equal frequency (median, quantile boundaries)



Discretization (bucketing)

Numerical data

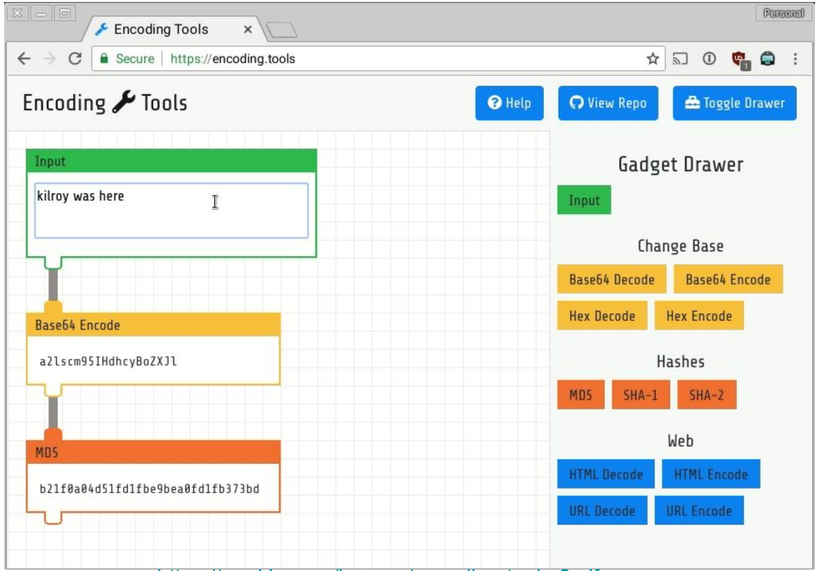
- When/why (not)?
  - May reduce the noise, which can improve a model's accuracy
    - Variable has more information than the problem needs
  - Prone to information loss

- Typical cases
  - Age, height, income

break

Data transformations
Categorical data

- Text data often needs to be converted to some type of numerical representation
    - Encoder



https://markhaa.se/images/encoding-tools-2.gif

Data transformations

Categorical data

- When/why
  - Many machine learning models do not accept categories as input
  - Attention: the model might interpret numerical values as weights

Data transformations

Categorical data

- Label encoding
  - Ordinal
    - With a column “First”, “Third”, and “Second” in it, the values can be directly mapped

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

- One-hot encoding
  - Binary columns from categories

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

- Feature hashing
  - Hash function to map categories to a smaller set of category columns
  - Avoid curse of dimensionality

Data transformations

Textual data

- Character normalization
  - In some languages, may need to convert all or some é è ë ê etc → e
  - In some others, may need to correct spelling Allén Allèn Allen → Allén
- Removing tags (HTML data)
  - If the data comes from a website, HTML tags can be eliminated (BeautifulSoup) leaving only the relevant content
- Bag-of-words representation
  - Ignore sentence structure and only consider documents
  - Document are word count vectors
    - For example, “This is a sentence with a sentence in it” → {“a”: 2, “in”: 1, “is”: 1, “it”: 1, “sentence”: 2, “this”: 1}
  - Term frequency–Inverse document frequency (TF-IDF)

Data transformations

Textual data

- Stemming and lemmatization
  - Word stems are usually the base form of possible words
  - **Stemming**: obtaining the base form of a word by removing prefix/suffix
    - WATCHES, WATCHING, and WATCHED = WATCH
  - **Lemmatization**: remove word affixes to get to the base form of a word, which is always a lexicographically correct word (root word)
  - Expanding contractions
    - wouldn't = would not; I'd = I would/had
- Removing stop-words
  - Words which have little or no significance in a text: a, the, and
  - Libraries provide standard stop-words lists for multiple languages

Data transformations

Feature selection and dimensionality reduction

- When/why
  - The higher the number of features, the harder it gets to visualize and explore the data
    - Select a subset of the original variables which captures as much information as the original set of variables
  - Often many of the features are highly correlated
  - It helps in data compression, reducing the memory and storage space needed
  - It helps reducing computation time
- Why not?
  - It may lead to some amount of data loss

Data transformations

Feature selection

- Feature selection techniques
  - Missing Value Ratio and Low Variance Filter
    - If a variable has too many missing values or always the same value, it will probably not have much added information
  - High Correlation Filter
    - Set a threshold (>= 0.5~0.6) to determine whether a highly correlated variable can be dropped
  - Feature importance
    - Some algorithms show the feature importance, and unimportant features can be discarded without impacting prediction capabilities
  - Backward/forward feature selection/elimination
    - All combinations from one to all variables *vs* model accuracy

Data transformations
Dimensionality reduction

- Dimensionality reduction algorithms
  - Single Value Decomposition (SVD)
    - SVD decomposes the original variables into three constituent matrices
      - S is the diagonal matrix of singular values, representing the importance values of different features in the matrix
    - Find the less significant features from the dataset that can be removed
    - Applications
      - Image Compression
      - Image Recovery
      - Spectral Clustering
      - Background Removal from Videos

Data transformations

Dimensionality reduction & feature selection

- Dimensionality reduction algorithms
  - Principal Component Analysis (PCA)
    - Extracts a new set of variables (principal components) from an existing large set of variables
    - A principal component is a linear combination of the original variables
    - Principal components are extracted in such a way that:
      - The first principal component explains maximum variance in the dataset
      - Second principal component tries to explain the remaining variance in the dataset and is uncorrelated to the first principal component
      - Third principal component tries to explain the variance which is not explained by the first two principal components and so on

Data transformations

Dimensionality reduction & feature selection

- Dimensionality reduction algorithms
  - Linear Discriminant Analysis (LDA)
    - Used when the data is labeled, unlike PCA and SVD
    - General approach is very similar to a PCA, but in addition to finding the component axes that maximize the variance of the data (PCA), LDA wants to maximize the separation between multiple classes
    - Dimensionality reduction does not only help reducing computational costs for a given classification task, but it can also be helpful to avoid overfitting by minimizing the error in parameter estimation (“curse of dimensionality”)