LO6

Data transformation for modelling

Objective

After attending this session, you should know

- Simple arithmetic
- Basic linear algebra
- Generating summary statistics
- Summarizing categorical data
- Parametric and non-parametric correlations analysis
- Transforming dataset distributions
- Extreme value and multivariate analysis for outliers

Arithmetic operator in python

- Array: a one-dimensional container for elements that are all of the same data type
- Matrix: a two-dimensional container for elements that are stored in an array
- The benefit of the NumPy library is that it makes it really easy to do math on data that's stored in either arrays or matrices.

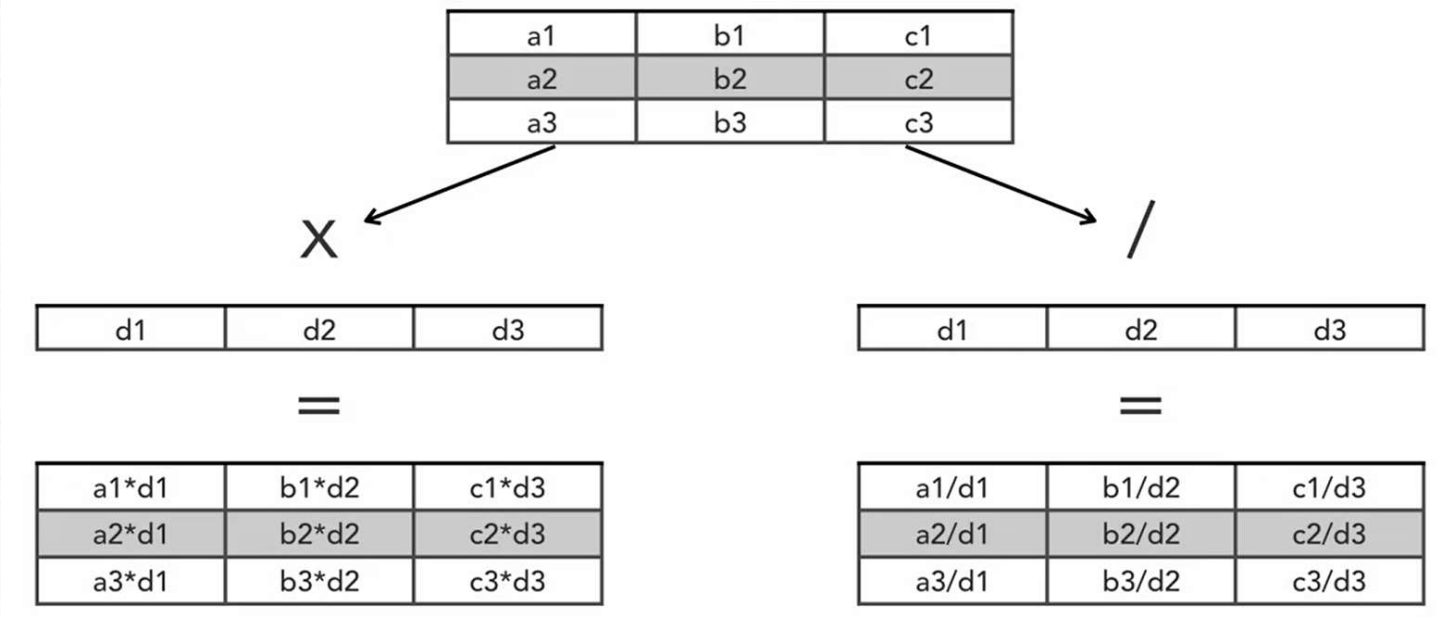
Class Exercises:

Time Series Manipulationhttps://www.geeksforgeeks.org/pandas-basic-of-time-series-manipulation/?ref=rp

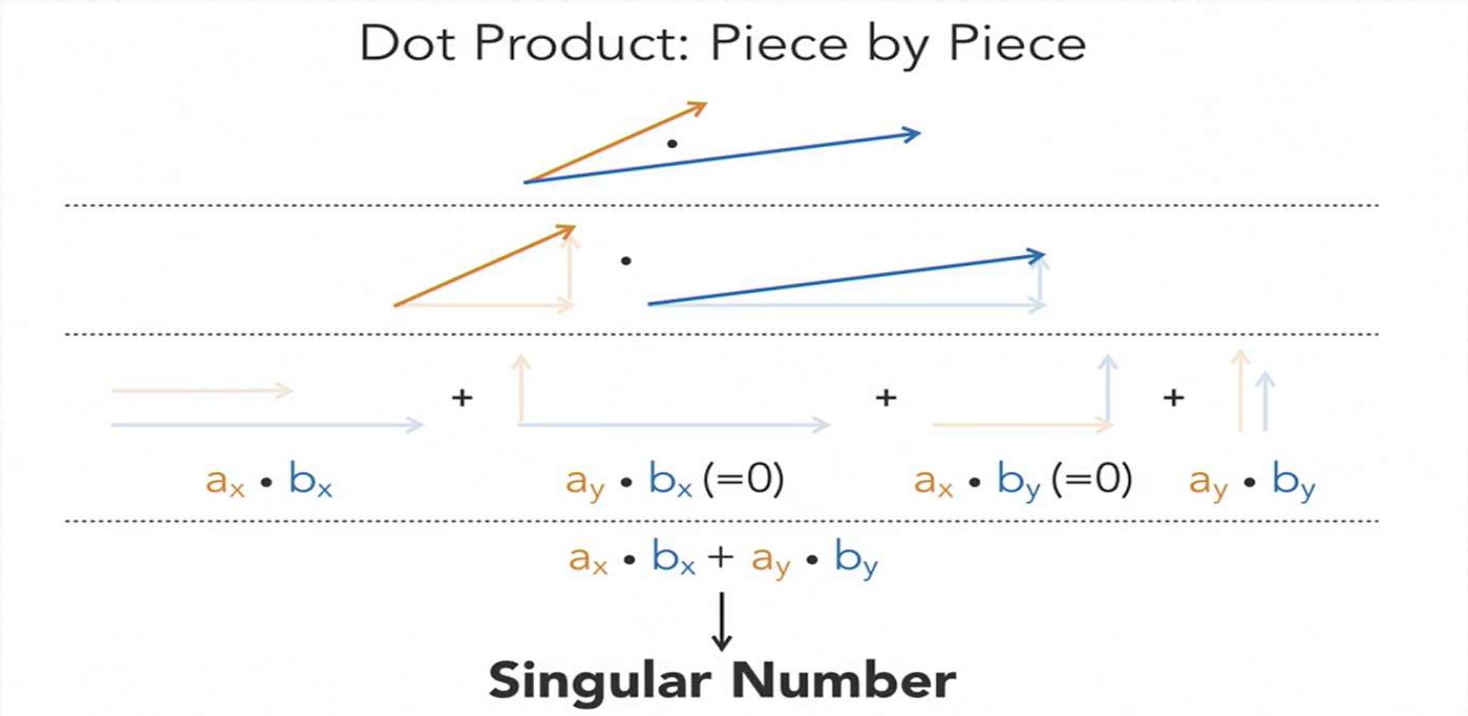
https://www.tutorialspoint.com/matrix-manipulation-in-python

Operator	Arithmetic Operation
+	Addition
-	Subtraction
*	Multiplication
/	Division

Arithmetic multiplication/division



Dot product


Dot Product: Piece by Piece
+
+
+
$a_x \cdot b_x$
$a_y \cdot b_x (=0)$
$a_x \cdot b_y (=0)$
$a_y \cdot b_y$
$a_x \cdot b_x + a_y \cdot b_y$
Singular Number

Matrix multiplication

2	4	6
1	3	5
10	20	30

•

0	1	2
3	4	5
6	7	8

2(0) + 4(3) + 6(6)	2(1) + 4(4) + 6(7)	2(2) + 4(5) + 6(8)
1(0) + 3(3) + 5(6)	1(1) + 3(4) + 5(7)	1(2) + 3(5) + 5(8)
10(0) + 20(3) + 30(6)	10(1) + 20(4) + 30(7)	10(2) + 20(5) + 30(8)

48	60	72
39	48	57
240	300	360

Descriptive statistics

- Descriptive statistics describe a variable's values and their spread.
- You can use them to get an understanding of a variable and the attributes that it represents.
- There are two categories of descriptive statistics.
  - describe the values of an observation in a variable
  - describe a variable's spread

Describe the observations in variable

- Sum
- Median
- Mean
- max

Describe variable spread

- Standard deviation
- Variance
- Counts
- quartiles

Use for Descriptive statistics

- Detecting outliers
- Planning data preparation requirement for data analysis
- Selecting the features for data analysis

Categorial variable

- Accepts only limited and fixed number values
- Each observation is assigned to a specific subgroup
- we could break down the dataset into apples and oranges based on this categorical fruits variable

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

Group
Apple
Orange

Creating a crosstab

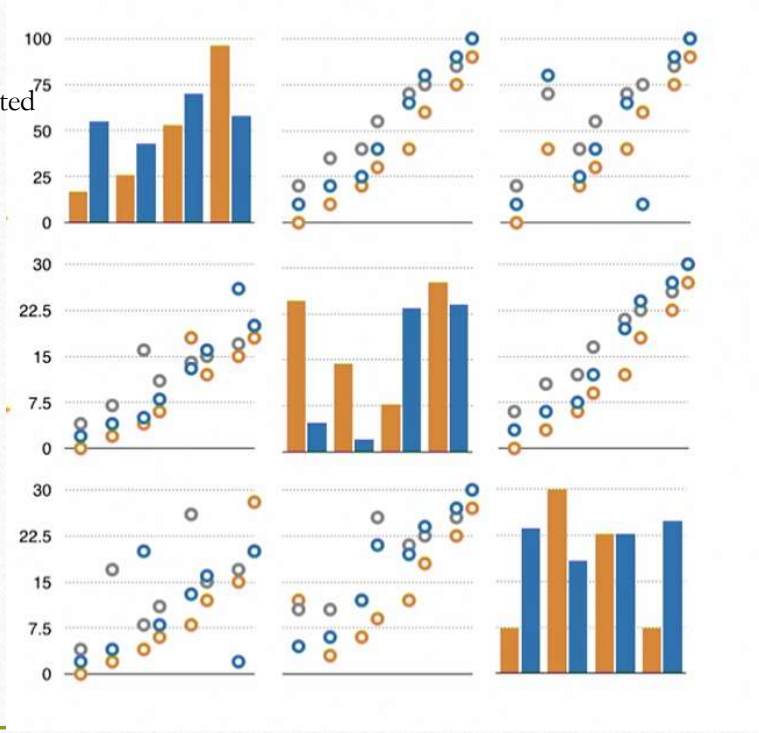
- A cross-tabulation of two or more features
- By default, a crosstab table shows frequency counts for features

	car_names	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gea	carb
car_names												
Mazda RX4	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4

Gear	3	4	5
am			
0	15	4	0
1	0	8	5

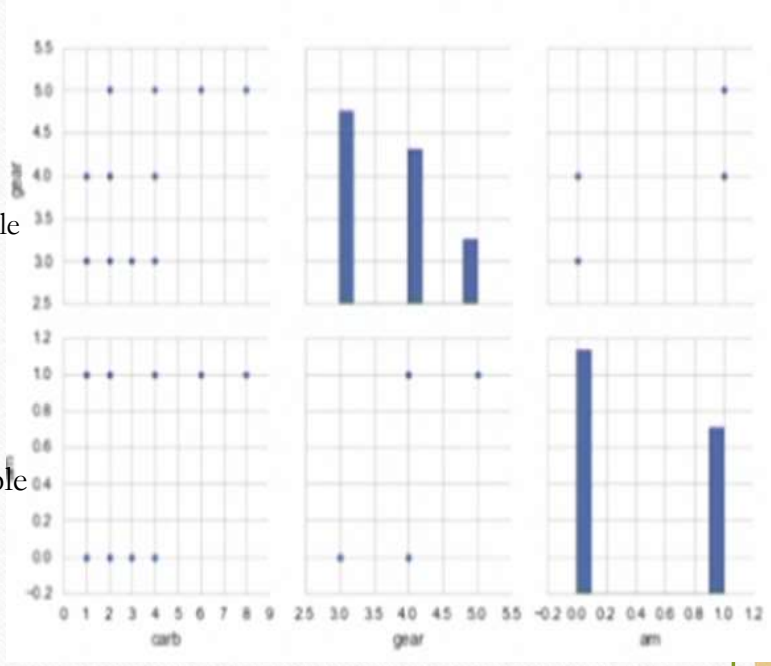
Normally distributed

Linearly related



Multinomial variable

Binomial variable



Parametric correlation analysis

- A method you can use to find correlation between linearly related continuous numeric variables
- **Pearson correlation coefficient**

  R = 1 (strong positive relationship)

  R = 0 (not linearly relationship)

  R = -1 (strong negative relationship)
- Your data is normally distributed
- You have continuous, numeric variables
- Your variables are linearly related

Use Pearson correlation

- To uncover (linear) relationship between variables
- Not to rule out possible (nonlinear) relationship between variables

Nonparametric correlation analysis

- You can use nonparametric correlation analysis to find correlation between categorical, nonlinearly related, non-normally distributed variables
- Spearman's rank correlation
- Chi-square tables
- For example, where nonparametric correlation analysis could be useful, imagine that you're a social scientist that studies smoking habits. You use a nonparametric correlation analysis, like Spearman's rank, to test the population for a correlation between income, as a bracket, and cigarette consumption of smokers. You find that higher income individuals are much more likely to smoke cigarettes than lower income people.

Spearman's rank correlation

- Find the R correlation between variable-pairs of ordinal data type (is a numeric variable that's able to be categorized)

  R = 1 (strong positive relationship)

  R = 0 (not correlated)

  R = -1 (strong negative relationship)

- Your variables are ordinal: numeric, but able to be ranked like a categorical variable
- Your variable are related nonlinearly
- Your variable are Non-normally distributed

Chi-square test for independence

- $P<0.05$

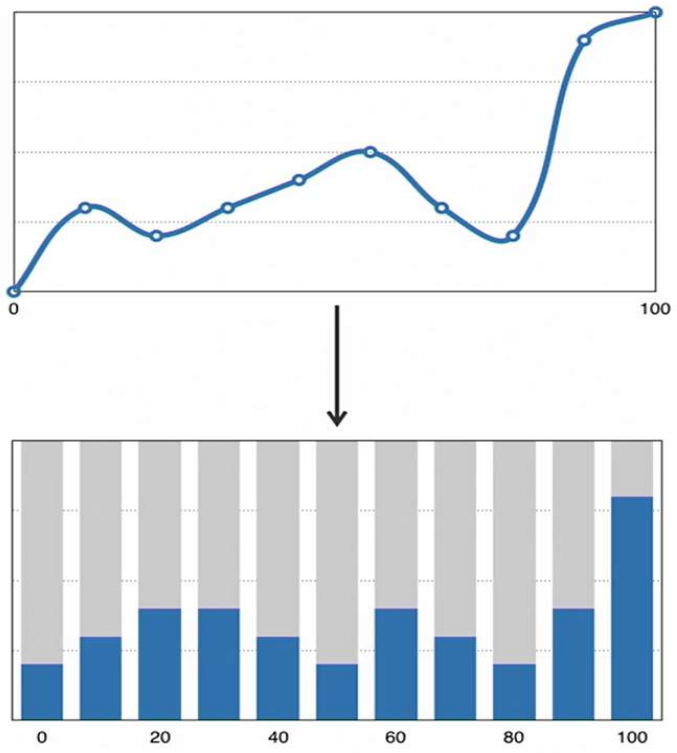
  Reject null hypothesis and conclude that the variables are correlated

- $P>0.05$

  accept null hypothesis and conclude that the variables are independent

Use chi-square table if

- Your variables are categorical or numeric
- You have binned the numeric variables
- Binning: imagine that you have a variable that had values between zero and 100. That's a numeric variable. As an example of binning, you could break up that variable into 10 separate groups, 10 groups of 10, and then within these 10 groups you would just put your data into different categories



Transforming dataset distributions

- For example, imagine you are in charge of sales and marketing for Zack's department store.
  - To measure the success of a recent holiday campaign you decide to compare daily sales revenues from a dataset in 1990 from one in 2016.
  - you measure the average sales revenues between November 15th and December 15th back in 1990. There is an average increase of $20 per checkout in this time period
  - but in 2016, the average increase was $200 per checkout. Is that net gain of $180 per checkout due to your marketing savvy?
  - No, it's due to other factors like monetary inflation and an increase in brand trust since 1990.
- You're trying to compare apples and oranges here because you forgot to scale your variables.

Why it's important to scale your data

- So that differing magnitude among variables don not produce erroneous or misleading statistics
- To prepare your data for data analysis / machine learning
- Two ways to scale your data

1. Normalization: putting each observation on a relative scale between value 0 and 1

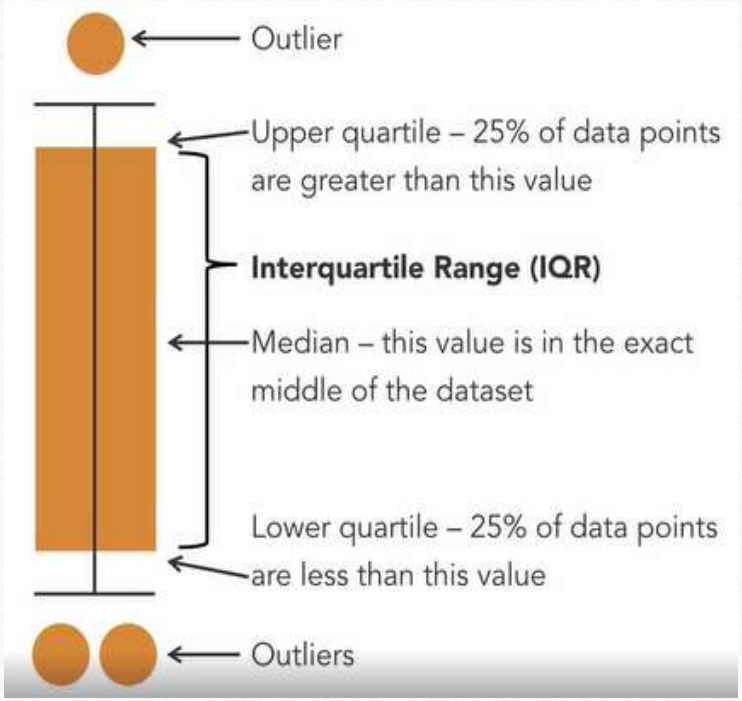
2. Standardization: rescaling data so that it has a zero mean and unit variance

Outlier detection

- Most machine learning methods assume that your data had been treated for outliers
- Detecting outliers can be a data preprocessing task or an analytical method of its own merit
- Use outlier detecting to uncover anomalies that represent:
  - Equipment failure
  - Fraud
  - Cybersecurity event

Univariate method: Tukey boxplots

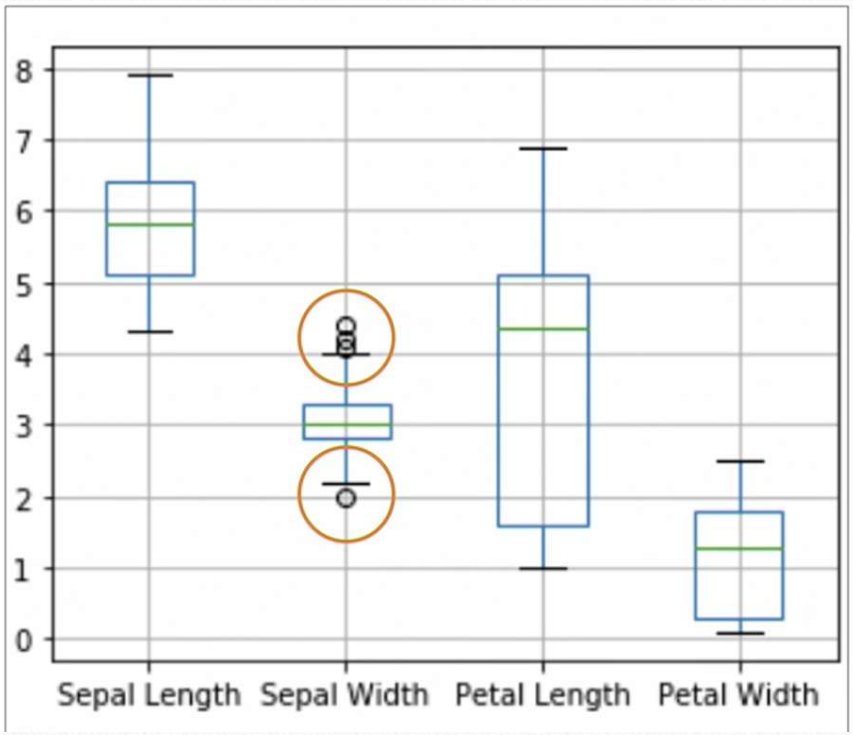
- Boxplot whiskers are set at 1.5*IQR
- If you see data points past these whiskers, they are outliers



Univariate method: Tukey outlier labelling

- a = Q1- 1.5(IQR)
- b = Q1+ 1.5(IQR)
- If…
- Min.value < a, or
- Max. value > b, then…
- The variable is suspect for outliers

Univariate method: Tukey boxplots



Sepal Width

> 4

< 2.5

Univariate method: Tukey boxplots

	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
15	5.7	4.4	1.5	0.4	setosa
32	5.2	4.1	1.5	0.1	setosa
33	5.5	4.2	1.4	0.2	setosa

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Species
60	5.0	2.0	3.5	1.0	versicolor

Univariate method: Tukey outlier labelling

IQR = 3.3 – 2.8 = 0.5

(1.5)IQR = 0.75

2.8 - 0.75 = 2.05

3.3 + 0.75 = 4.05

	0	1	2	3
count	150.0	150.0	150.0	150.0
mean	5.8	3.1	3.8	1.2
std	0.8	0.4	1.8	0.8
min	4.3	2.0	1.0	0.1
25%	5.1	2.8	1.6	0.3
50%	5.8	3.0	4.3	1.3
75%	6.4	3.3	5.1	1.8
max	7.9	4.4	6.9	2.5

Multivariate outlier detection

- Use multivariate methods to find the outliers that only show up within combinations of observations from two or more different variables
- There are many different multivariate methods to detect outliers

Summary

- Simple arithmetic
- Basic linear algebra
- Generating summary statistics
- Summarizing categorical data
- Parametric and non-parametric correlations analysis
- Transforming dataset distributions
- Extreme value and multivariate analysis for outliers

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