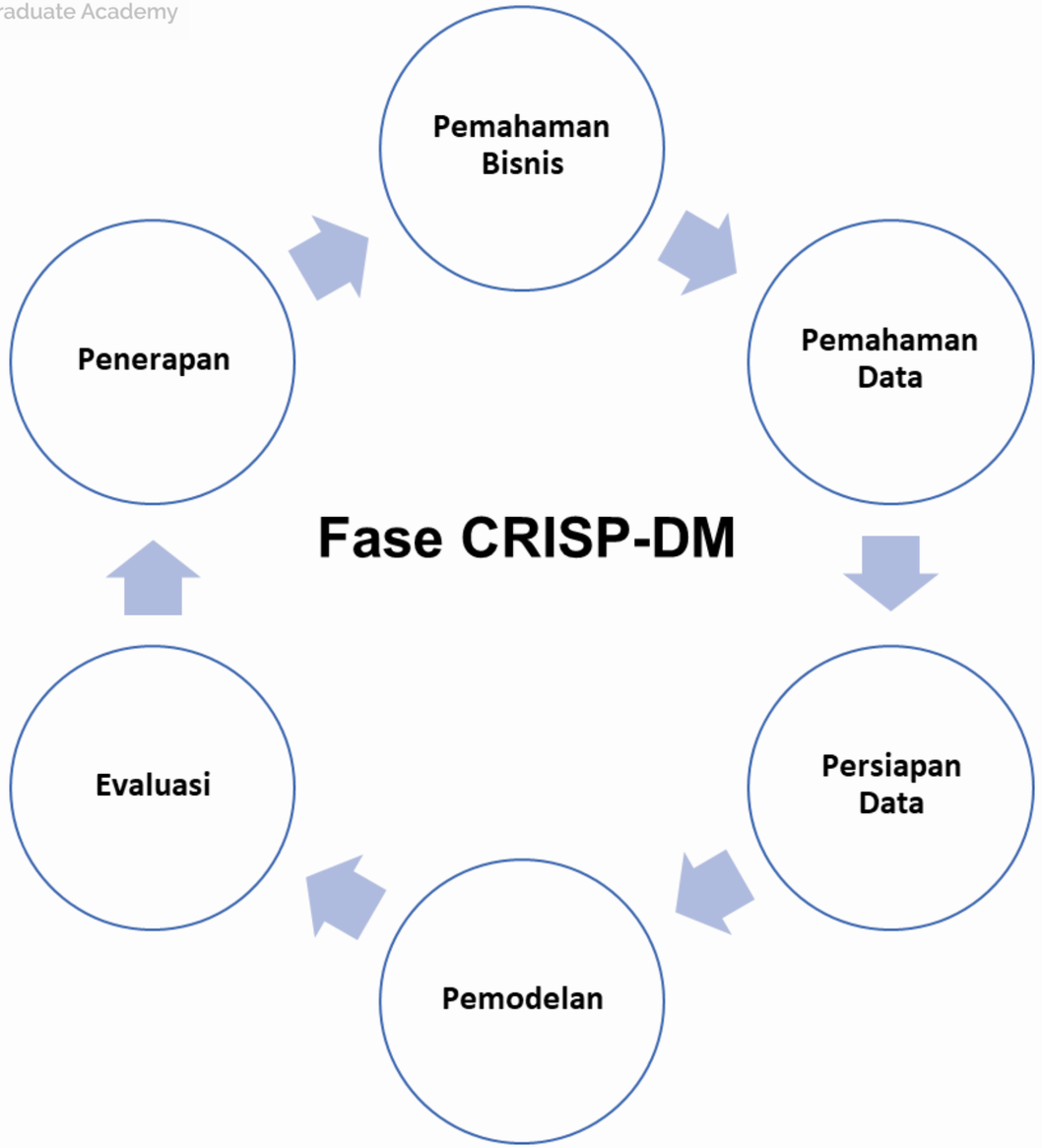
Pemilihan dan Pembersihan Data



PEMILIHAN DATA

Pemilihan Data

- Sebelumnya pada tahap data understanding sudah dilakukan proses pengumpulan data
- Selanjutnya dilakukan pemilihan data yang relevan dengan tujuan data mining yang telah ditentukan.
- Secara umum, terdapat 2 cara pemilihan data:
  - Memilih item (baris)
  - Memilih atribut (kolom)

```
#delete the cabin feature/column and others previously stated to exclude in train data
drop_column = ['PassengerId','Cabin', 'Ticket']
data1.drop(drop_column, axis=1, inplace = True)
```

Data Cleaning
(Pembersihan Data)

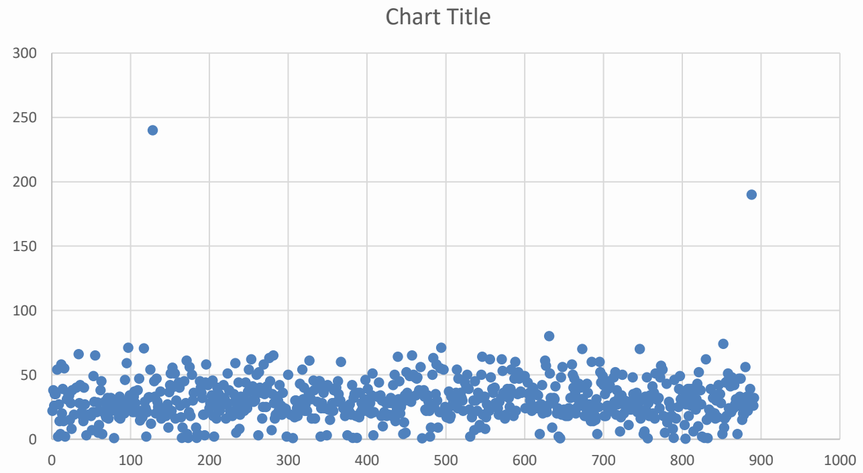
Data Cleaning

- Data cleaning (pembersihan data) merupakan tindak lanjut dari hasil analisa kualitas data pada tahap Data Understanding

Masalah	Solusi
Missing data	Exclude baris/kolom atau isi dengan nilai perkiraan
Data error	Exclude baris/kolom atau gunakan logika untuk memperbaiki nilai
Coding inconsistencies	Pilih salah satu coding scheme kemudian sesuaikan data
Missing/bad metadata	Cek secara manual kemudian tentukan definisi yang benar

Data Cleaning – Correcting

- Correcting → melakukan review data, jika menemukan nilai yang unik pada satu diantara yang lainnya, maka perlu untuk dilakukan perbaikan pada data tersebut.
- Misal pada data Age = 800, maka perlu diperbaiki menjadi 80. Jika ragu dengan nilai yang benar, maka data tersebut lebih baik diexclude dari dataset



Data Cleaning – Completing

- Tentukan kolom/fitur yang diyakini akan mempengaruhi hasil data mining. Kolom/fitur yang tidak penting bisa diexclude
- Nilai Null atau data yang hilang harus dilengkapi. Hal tersebut dilakukan karena beberapa algoritma (ex: NN Classification) tidak bisa menghandle nilai null, dan akan terjadi error di iterasi awal.
- Dua cara yang dilakukan adalah dengan menghapus record atau mengisikan missing values dengan data yang dapat dipertanggungjawabkan.

Data Cleaning – Completing (2)

- Menghapus data yang memiliki missing values sangat tidak disarankan, apalagi jika jumlah data/record bermissing values persentasenya besar, kecuali jika record tersebut tidak mungkin dilengkapi.
- Mengisikan missing values adalah pilihan terbaik karena satu record bisa sangat mempengaruhi hasil data mining.
- Metodologi dasar untuk mengisikan missing values pada variabel kualitatif adalah menggunakan modus, sedangkan pada variabel kuantitatif menggunakan mean, median, atau mean+standard deviasi acak.
- Metodologi lain yang digunakan adalah dengan menggunakan kriteria khusus. Contoh: usia rata-rata pada Class tertentu

Data Cleaning – Converting

- Converting: melakukan konversi format tipe data.
- Biasanya dilakukan encoding data string ke number categorical.
- Contoh Sex_code (M/F) – diencode menjadi (0/1). Hal ini dilakukan untuk mempermudah perhitungan median/modus.

Data Cleaning – Creating

- Creating: disebut juga dengan Feature Engineering yaitu menggunakan fitur yang ada untuk membuat fitur baru yang kira-kira dapat meningkatkan performa data mining.
- Proses creating data akan dibahas pada pertemuan selanjutnya

```
for dataset in data_cleaner:
    #complete missing age with median
    dataset['Age'].fillna(dataset['Age'].median(), inplace = True)

    #complete embarked with mode
    dataset['Embarked'].fillna(dataset['Embarked'].mode()[0], inplace = True)

    #complete missing fare with median
    dataset['Fare'].fillna(dataset['Fare'].median(), inplace = True)
```

```
    dataset['FamilySize'] = dataset ['SibSp'] + dataset['Parch'] + 1

    dataset['IsAlone'] = 1 #initialize to yes/1 is alone
    dataset['IsAlone'].loc[dataset['FamilySize'] > 1] = 0 # now update to no/0 if family size is greater than 1

    #quick and dirty code split title from name: http://www.pythonforbeginners.com/dictionary/python-split
    dataset['Title'] = dataset['Name'].str.split(", ", expand=True)[1].str.split(".", expand=True)[0]
```

REVIEW
Eksplorasi Data

Descriptive Statistics

- Descriptive Statistics adalah pengukuran statistik yang bertujuan melihat kesimpulan fitur2 penting pada dataset.
- Beberapa contoh:
  - Measures of Center
  - Measures of Spread

Measures of Center

- Measures of Center adalah descriptive statistics yang melihat “ketengahan” dari data numeric, misalnya dengan menentukan nilai Mean, Median, dan Modus.

Measures of Center

In [1]:
```
%matplotlib inline

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
mtcars = pd.read_csv("../input/mtcars/mtcars.csv")
mtcars =  mtcars.rename(columns={'Unnamed: 0': 'model'})
mtcars.index = mtcars.model
del mtcars["model"]

mtcars.head()
```

Measures of Center

Out[2]:

model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2

Mean

In [3]:
```
mtcars.mean()                    # Get the mean of each column
```

Out[3]:
```
mpg      20.090625
cyl       6.187500
disp    230.721875
hp      146.687500
drat      3.596563
wt        3.217250
qsec     17.848750
vs        0.437500
am        0.406250
gear      3.687500
carb      2.812500
dtype: float64
```

Mean of Row

In [4]:
```
mtcars.mean(axis=1)          # Get the mean of each row
```

Out[4]:

model	
Mazda RX4	29.907273
Mazda RX4 Wag	29.981364
Datsun 710	23.598182
Hornet 4 Drive	38.739545
Hornet Sportabout	53.664545
Valiant	35.049091
Duster 360	59.720000
Merc 240D	24.634545
Merc 230	27.233636
Merc 280	31.860000
Merc 280C	31.787273
Merc 450SE	46.430909
Merc 450SL	46.500000
Merc 450SLC	46.350000
Cadillac Fleetwood	66.232727
Lincoln Continental	66.058545
Chrysler Imperial	65.972273
Fiat 128	19.440909
Honda Civic	17.742273

Median

In [5]:
```
mtcars.median()                    # Get the median of each column
```

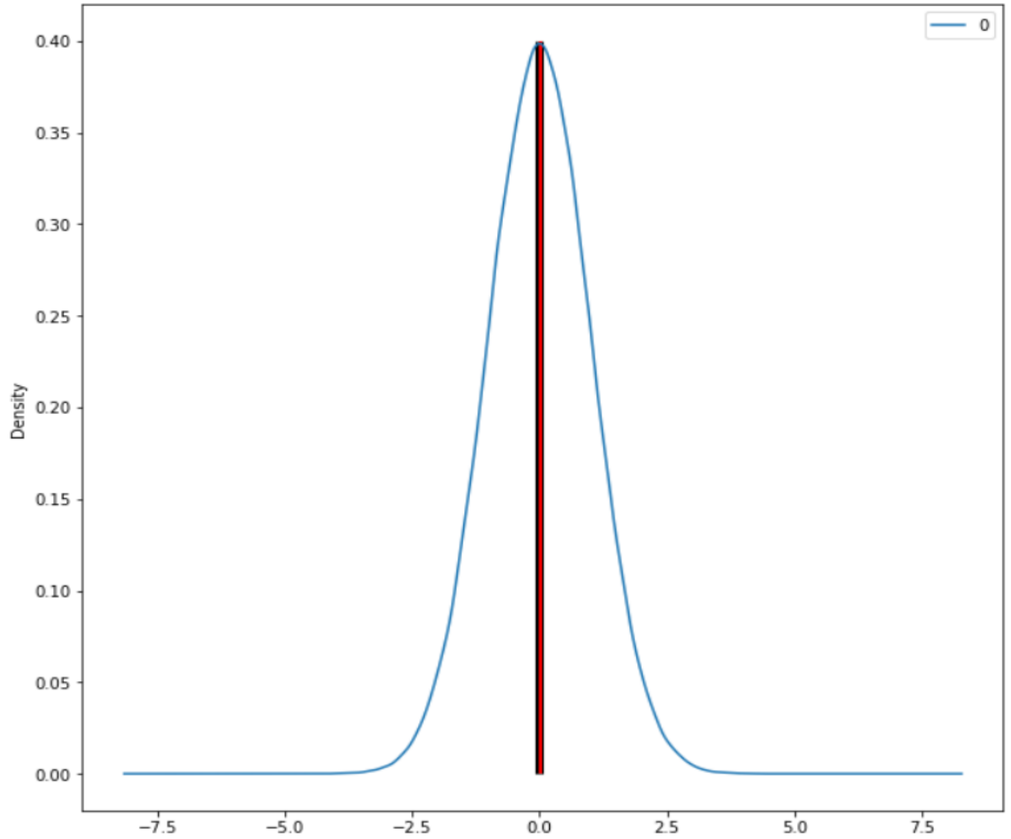
Out[5]:
```
mpg      19.200
cyl       6.000
disp    196.300
hp      123.000
drat      3.695
wt        3.325
qsec     17.710
vs        0.000
am        0.000
gear      4.000
carb      2.000
dtype: float64
```

Measures of Center

- Mean dan Median digunakan untuk mengidentifikasi pusat data, tetapi dua descriptive statistic ini tidak selalu memberi nilai yang sama.
- Median selalu memberi nilai yang membagi data menjadi dua bagian sama banyak, mean memberikan nilai rata-rata numerik sehingga nilai extreme dapat memberikan hasil selisih yang signifikan pada mean.
- Pada distribusi simetris, mean dan median akan menghasilkan nilai yang sama.

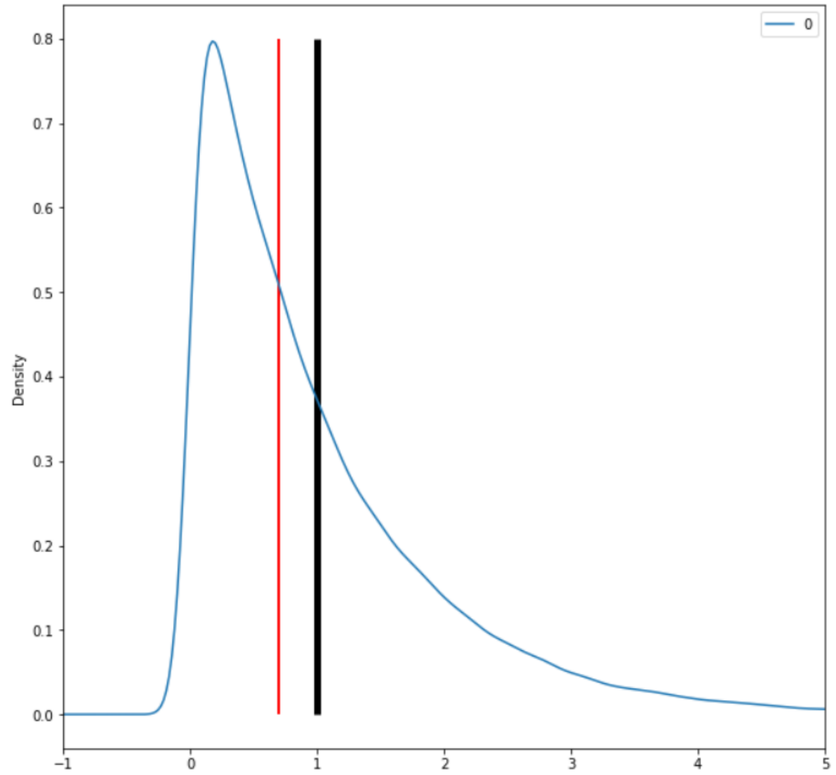
Measures of Center

- Pada distribusi simetris, mean dan median akan menghasilkan nilai yang sama.



Measures of Center

- Pada data skew, mean akan tertarik ke bagian yg melandai (skewed), sedangkan median tidak terlalu berpengaruh pada effect skewed



Measures of Center

- Mean sangat dipengaruhi oleh data anomali (outliers), sedangkan median lebih tahan terhadap data outliers
- Karena median lebih tahan terhadap skewness dan outliers, median dikenal dengan nama “robust” stats. Median menghasilkan nilai yg lebih akurat pada distribusi data yg memiliki skew dan outliers.

Modus

- Modus menghasilkan nilai pada data yang paling sering muncul.
- Tidak seperti mean dan median, pada modus kita bisa mendapatkan beberapa hasil modus (multiple modes).
- Pada contoh diatas terdapat beberapa kolom dengan nilai modus lebih dari satu. Yang memiliki data tidak lebih dari satu maka akan dituliskan NaN

Out[2]:

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
model											
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2

Modus

```
mtcars.mode()
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	10.4	8.0	275.8	110.0	3.07	3.44	17.02	0.0	0.0	3.0	2.0
1	15.2	NaN	NaN	175.0	3.92	NaN	18.90	NaN	NaN	NaN	4.0
2	19.2	NaN	NaN	180.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	21.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	21.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	22.8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	30.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Measures of Spread

- Measures of Spread (dispersi) adalah descriptive statistics yang menunjukkan sebaran data.
- MoS yang paling sederhana adalah range. Range adalah jarak antara nilai maximum dan nilai minimum

```
max(mtcars["mpg"]) - min(mtcars["mpg"])
```

```
23.5
```

Quantile

- Quantiles adalah nilai yang membagi data terurut menjadi beberapa bagian dengan jumlah data yang sama
- Quantile juga biasa digunakan untuk melihat sebaran data.

```
five_num = [mtcars["mpg"].quantile(0),
            mtcars["mpg"].quantile(0.25),
            mtcars["mpg"].quantile(0.50),
            mtcars["mpg"].quantile(0.75),
            mtcars["mpg"].quantile(1)]

five_num
```

```
[10.4, 15.425, 19.2, 22.8, 33.9]
```

```
mtcars["mpg"].describe()
```

```
count    32.000000
mean     20.090625
std       6.026948
min      10.400000
25%      15.425000
50%      19.200000
75%      22.800000
max      33.900000
Name: mpg, dtype: float64
```

Varians & Standar Deviasi

- Varians dan Standar Deviasi juga umum digunakan untuk melihat sebaran data. Standar Deviasi merupakan akar pangkat dua dari varians

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

- Standar deviasi mengukur penyebaran kelompok data terhadap nilai mean.
- Jika nilai standar deviasi tinggi maka sebaran data luas

Varians & Standar Deviasi

```
mtcars["mpg"].var()
```

```
36.32410282258065
```

```
mtcars["mpg"].std()
```

```
6.026948052089105
```