

Big Data Analysis Data preparation and cleansing:

Kohei Arai

1

Big Data Analysis

- Data analysis methods are classified by purpose and introduced according to the analysis process.
- The flow from data collection and preparation to advanced analysis, and then to decision-making and action.
- By appropriately combining these analysis methods, data-based decision-making and strategy planning become possible.

2

Python

- Basic Python code example (simple numerical calculation example)
- Define a function called add_numbers and call it to add up some numbers
- The result is displayed in the console

```
# Function to add two numbers
def add_numbers(a, b):
    result = a + b
    return result
# Call function
num1 = 5
num2 = 10
sum_result =
add_numbers(num1, num2)
# Display result
print(f"The sum of {num1}
and {num2} is: {sum_result}")
```

3

1. The first step in data analysis is to ensure high-quality data.

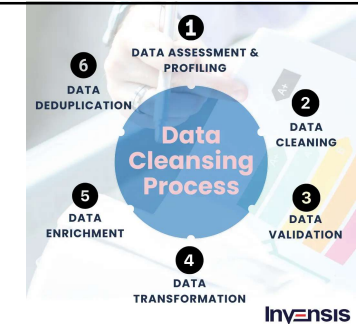
- Data Cleansing/Data Cleaning: The process of preparing data for analysis, including handling missing values, removing outliers, and standardizing data formats ⇒ the foundation for obtaining accurate analysis results
- Exploratory Data Analysis: An initial analysis method for understanding the characteristics of data, including checking basic statistics of data, visualizing distribution, and grasping trends

4

Data Cleansing

- Data cleansing is the process of removing or correcting inaccurate, incomplete, or invalid data from a dataset to improve data quality. Simple Data Cleansing Example Using Python
- Perform basic data cleansing techniques such as removing missing values, converting to numerical data, and removing special characters in strings.
- Data cleansing varies depending on the project and data set, so it is important to select and customize the method based on the specific requirements.
- In actual work, various data cleansing operations are required according to the characteristics of the data set, such as missing value processing, outlier detection and correction, abnormal value completion, and data type conversion.

5



6

```
import pandas as pd
# Read CSV file
df = pd.read_csv('data.csv')
# Display the first 5 rows
print(df.head())
# Check for missing values
print(df.isnull().sum())
# Remove rows with missing values
df = df.dropna()
# Convert 'age' column to numerical
df['age'] = df['age'].astype(int)
# Remove rows with non-numeric ages
df = df[df['age'].notnull()]
# Standardize 'salary' column
df['salary'] = df['salary'].str.replace('$', '').str.replace(',', '').astype(float)
# Remove rows with non-numeric salaries
df = df[df['salary'].notnull()]
# Display the first 5 rows after cleansing
print(df.head())
```

7

Data cleaning

- Data cleaning is the process of detecting and correcting errors and inaccurate values in a dataset to improve data quality. Examples of basic data cleaning operations using Python's pandas library.
- Perform basic data cleaning techniques such as removing missing values, correcting negative ages, correcting inaccurate gender, and converting non-numeric salaries to missing values.
- It is important to select and customize the appropriate cleaning technique based on the characteristics of the dataset and business requirements.
- In real-world work, data understanding and domain knowledge are essential, and data cleaning is performed carefully and continuously by data scientists and data engineers.

8



9

10

Exploratory Data Analysis

11

12

13

14

15

16

17

18

```
print(data.isnull().sum())
# ヒストグラムの描画
data.hist(figsize=(12, 10))
plt.show()
# 相関行列のヒートマップ
correlation_matrix = data.corr()
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True,
            cmap='coolwarm', fmt='.2f')
plt.show()
# 散布図行列
sns.pairplot(data)
plt.show()
```

19

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 12 columns):
#   Column      Non-Null Count  Dtype
---  -
0   age          30 non-null     int64
1   income       30 non-null     int64
2   education years 30 non-null     int64
3   work experience 30 non-null     int64
4   credit score  30 non-null     int64
5   debt to income 30 non-null     float64
6   has children  30 non-null     int64
7   home ownership 30 non-null     int64
8   loan amount  30 non-null     int64
9   interest rate 30 non-null     float64
10  loan term     30 non-null     int64
11  loan status   30 non-null     int64
dtypes: float64(2), int64(10)
memory usage: 2.9 KB
None
```

20

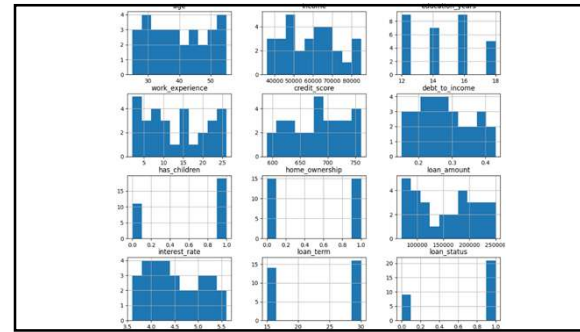
	age	income	education years	work experience	credit score	¥
count	30.000000	30.00000	30.000000	30.000000	30.000000	30.000000
mean	39.400000	58500.00000	14.666667	13.000000	681.000000	681.000000
std	9.360961	14143.96423	2.186689	7.799204	49.626189	49.626189
min	25.000000	36000.00000	12.000000	2.000000	590.000000	590.000000
25%	31.250000	47250.00000	12.000000	6.250000	642.500000	642.500000
50%	38.500000	58500.00000	14.000000	11.500000	685.000000	685.000000
75%	47.500000	67750.00000	16.000000	19.750000	717.500000	717.500000
max	55.000000	85000.00000	18.000000	26.000000	760.000000	760.000000

	debt to income	has children	home ownership	loan amount	¥
count	30.000000	30.000000	30.000000	30.000000	30.000000
mean	0.278000	0.633333	0.500000	154833.333333	154833.333333
std	0.079498	0.490133	0.508548	58464.302892	58464.302892
min	0.150000	0.000000	0.000000	70000.000000	70000.000000
25%	0.220000	0.000000	0.000000	96250.000000	96250.000000
50%	0.275000	1.000000	0.500000	155000.000000	155000.000000
75%	0.327500	1.000000	1.000000	200000.000000	200000.000000
max	0.430000	1.000000	1.000000	250000.000000	250000.000000

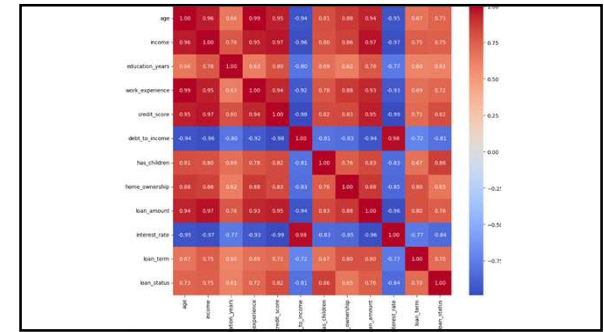
21

```
interest rate  loan term  loan status
count      30.000000    30.000000    30.000000
mean      4.466667    23.000000    0.700000
std       0.588003    7.611244    0.466092
min       3.600000    15.000000    0.000000
25%      4.000000    15.000000    0.000000
50%      4.350000    30.000000    1.000000
75%      4.950000    30.000000    1.000000
max       5.600000    30.000000    1.000000
age         0
income      0
education years 0
work experience 0
credit score 0
debt to income 0
has children 0
home ownership 0
loan amount 0
interest rate 0
loan term   0
loan status 0
dtype: int64
```

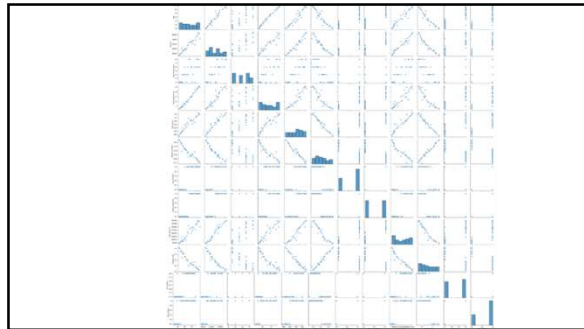
22



23



24



25