

Class period 10

บทที่ 5 การวิเคราะห์ข้อมูลสำหรับการตัดสินใจ 2 (ต่อ)
Pandas 102 part2

1

.groupby()

- <https://www.kaggle.com/code/zenfroid/python-groupby-tutorial>
- วิธีการแยกข้อมูลตามเงื่อนไขคือการ โดยนำค่าที่อยู่ในเงื่อนไขมาแยกเป็นกลุ่มตามเงื่อนไขที่กำหนด เช่น
- `data_covid.groupby('nationality')`
- ตัวอย่างในการใช้งานด้วย groupby (ใช้ค่าในเงื่อนไขที่กำหนดให้ไปแยกข้อมูลตามเงื่อนไขที่กำหนดให้) เช่น list เช่น ["จีน", "อินเดีย"]
- จากตัวอย่างการแยกข้อมูลตามเงื่อนไข nationality ค่านี้ groupby จะทำการแยกข้อมูลตามกลุ่มของ nationality โดยที่กลุ่มตามค่าที่อยู่ในเงื่อนไขเป็น nationality
- groupby จะทำซ้ำ memory การประมวลผลจากการใช้ groupby จะทำให้ได้ผลลัพธ์ตามความต้องการในการคำนวณค่าที่ต้องการดูจาก data_covid.groupby('nationality').count()

2

คำสั่งที่ใช้สำหรับผลลัพธ์ของ .groupby()

- ค่าตัวอย่าง
- count() ใช้สำหรับดูจำนวนค่าในแต่ละคอลัมน์ในแต่ละกลุ่มตามค่าที่อยู่ในเงื่อนไขที่กำหนดในคอลัมน์ที่ใช้ groupby
- max() ใช้ดูค่าค่าที่มากที่สุดในแต่ละกลุ่ม (ดูในค่าคอลัมน์ที่มีค่าอยู่ในวงเล็บ)
- min() ใช้ดูค่าค่าที่น้อยที่สุดในแต่ละกลุ่ม (ดูในค่าคอลัมน์ที่มีค่าอยู่ในวงเล็บ)

Arithmetic operators	String operations	Math comparison operators
+	+	>
-	-	<
*	*	>=
/	/	<=
%	%	==
**	**	!=
^	^	>>
~	~	<<
~	~	<<
~	~	<<
~	~	<<

3

เฉลย Homework class period 9 ด้วย groupby()

- สร้างตารางใหม่ ขึ้นมาใน sex เป็น missing ที่ประเทศ
- `data_covid['sex'].isnull()`
- ตรวจสอบค่าที่ missing ในคอลัมน์ sex และสร้าง list logical expression True(missing)/False(not missing)
- `missing_sex = data_covid[data_covid['sex'].isnull()]`
- ใช้ list logical expression มาใช้กับข้อมูลในตาราง records ที่มีในคอลัมน์ sex เป็น missing และนำค่ามาที่ติดกับใน sex เป็น missing_sex
- `missing_sex` แสดงค่าตาราง records ที่มีในคอลัมน์ sex เป็น missing

4

เฉลย Homework class period 9 ด้วย groupby()

- สร้างตารางใหม่ ขึ้นมาใน sex เป็น missing ที่ประเทศ
- `data_covid['sex'].isnull()`
- ตรวจสอบค่าที่ missing ในคอลัมน์ sex และสร้าง list logical expression True(missing)/False(not missing)
- `missing_sex = data_covid[data_covid['sex'].isnull()]`
- ใช้ list logical expression มาใช้กับข้อมูลในตาราง records ที่มีในคอลัมน์ sex เป็น missing และนำค่ามาที่ติดกับใน sex เป็น missing_sex
- `missing_sex` แสดงค่าตาราง records ที่มีในคอลัมน์ sex เป็น missing

5

create pandas table

Dictionary	List
<pre>records = [{"account": "Jenex LLC", "Jan": 150, "Feb": 200, "Mar": 140}, {"account": "Alpha Co", "Jan": 200, "Feb": 210, "Mar": 215}, {"account": "Blue Inc", "Jan": 180, "Feb": 190, "Mar": 195}]</pre>	<pre>records = [{"account": "Jenex LLC", "Jan": 150, "Feb": 200, "Mar": 140}, {"account": "Alpha Co", "Jan": 200, "Feb": 210, "Mar": 215}, {"account": "Blue Inc", "Jan": 180, "Feb": 190, "Mar": 195}]</pre>

6

ตัวอย่างการสร้างตาราง pandas

- แบบ Dictionary ให้ pd.DataFrame()
- ขั้นตอนการสร้างตาราง ขึ้นมาและเขียนค่าแต่ละ record ที่ต้องการในรูปของ dictionary โดย index จะเป็นชื่อคอลัมน์และ value จะเป็นค่าของ record นั้นๆ เช่น
- `records = [{"account": "Jenex LLC", "Jan": 150, "Feb": 200, "Mar": 140}, {"account": "Alpha Co", "Jan": 200, "Feb": 210, "Mar": 215}, {"account": "Blue Inc", "Jan": 180, "Feb": 190, "Mar": 195}]`
- `records_df = pd.DataFrame(records)`
- `records_df`

7

ตัวอย่างการสร้างตาราง pandas

- แบบ List ให้ pd.DataFrame.from_records()
- ขั้นตอนการสร้างตาราง ขึ้นมาและเขียนค่าแต่ละ record ที่ต้องการในรูปของ dictionary โดย index จะเป็นชื่อคอลัมน์และ value จะเป็นค่าของ record นั้นๆ เช่น
- `records = [{"account": "Jenex LLC", "Jan": 150, "Feb": 200, "Mar": 140}, {"account": "Alpha Co", "Jan": 200, "Feb": 210, "Mar": 215}, {"account": "Blue Inc", "Jan": 180, "Feb": 190, "Mar": 195}]`
- `records_df = pd.DataFrame.from_records(records)`
- `records_df`

8

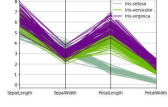
Simple Visualization

- โฉมหน้ากราฟของข้อมูลตามนี้คือ 3x3 จากนี้คือ
- <https://www.kaggle.com/pandas-dev/pandas-visualization/pandas-visualization/data-csv/viz.csv>
- ดาวน์โหลดข้อมูลจาก link และนำข้อมูลไปใช้กับ
- `df = pd.read_csv("https://www.kaggle.com/pandas-dev/pandas-visualization/pandas-visualization/data-csv/viz.csv")`
- `df`
- ผลลัพธ์ของ `df.groupby('Name').count()` ดูที่ตัวอย่างนี้

9

parallel_coordinates

- `pd.plotting.parallel_coordinates(df, "Name")`
- การกำหนดให้ชื่อเส้นเป็นนามสกุล
- แต่ละเส้นแทน record เป็นนามสกุล y
- โดยแต่ละเส้นแสดงค่าของ feature เป็น 1 จุด 1 record คือ 1 เส้น อาจจะมีหลายเส้นแสดงค่าของ record เป็นค่าในชุดของเส้น
- ใช้ดูแนวโน้มการเปลี่ยนแปลงของ feature



9

10

scatter_matrix

- `pd.plotting.scatter_matrix(df)`
- การนำค่าที่เป็นตัวเลขมาแสดง record ในแต่ละคอลัมน์นำมาเขียนเป็นเส้น
- ตัวเป็นนามสกุล x และ y จะเหมือนกัน คือชื่อของเส้น
- แต่ละเส้นแสดงค่าของ feature เป็น 1 จุด 1 record คือ 1 เส้น อาจจะมีหลายเส้นแสดงค่าของ record เป็นค่าในชุดของเส้น
- ใช้ดูแนวโน้มการเปลี่ยนแปลงของ feature
- สามารถใช้ดูความสัมพันธ์ของ feature ได้




10

11

save table

- ใช้คำสั่ง `to_csv()` ในการบันทึกเป็นไฟล์ csv
- `missing_sex = data_covid[data_covid['sex'].isnull()]`
- `missing_sex`
- การนำข้อมูลที่ขาดหายไปมาแสดง
- ใช้คำสั่ง `missing_sex.to_csv('missing_sex.csv')`
- `missing_sex.to_csv('missing_sex.csv')`



11

12