ASIAN INSTITUTE OF MANAGEMENT

Data Mining and Wrangling

The Prologue

Session 1

BSDSBA 2028

21 January 2026

ASIAN INSTITUTE OF MANAGEMENT

Session 1 – The Prologue

Gameplan

11:00 AM to 11:30 AM	The Class
11:30 AM to 12:00 NN	What is DMW?
12:00 NN to 12:30 PM	First Class Activity
12:30 PM to 1:00 PM	**Break**

Prologue I
The Class

Course Learning Outcomes

1 Explain the different procedures in data wrangling and mining for various data types.

2 Collect and mine data from various data sources using various techniques.

3 Generate hypotheses and derive insights from different forms of datasets by having operational knowhow in data mining and wrangling.

4 Write and present technical reports on data analysis for a specialized audience.

Grading Criteria and Course Deliverables

- **In-Class Activities – 5%**
- **Class Participation – 10%**
  - Attendance – 60%
  - Post-class Reflection – 20%
  - Class Contribution (Recitation/Discussion Boards) – 20%
- **Exercises – 15%**
- **Assignments – 15%**
- **Mini-Projects (Lab Reports) – 15%**
- **Final Project Report and Presentation – 20%**
- **Midterm and Final Examination – 20%**

Generative AI Policy

Generative AI technology is considered a **tool or reference** similar to Wikipedia and Stack Overflow. You may use them in the same manner as you use these tools, and just like them, should be **cited and acknowledged**. Your submissions should be your intellectual output and **not lifted directly** from their output. Passing off their output as your own will be considered as an **academic misconduct** and will be reported to the Office of the Dean for appropriate action.

Prologue II
What is DMW?

What is Data Mining?

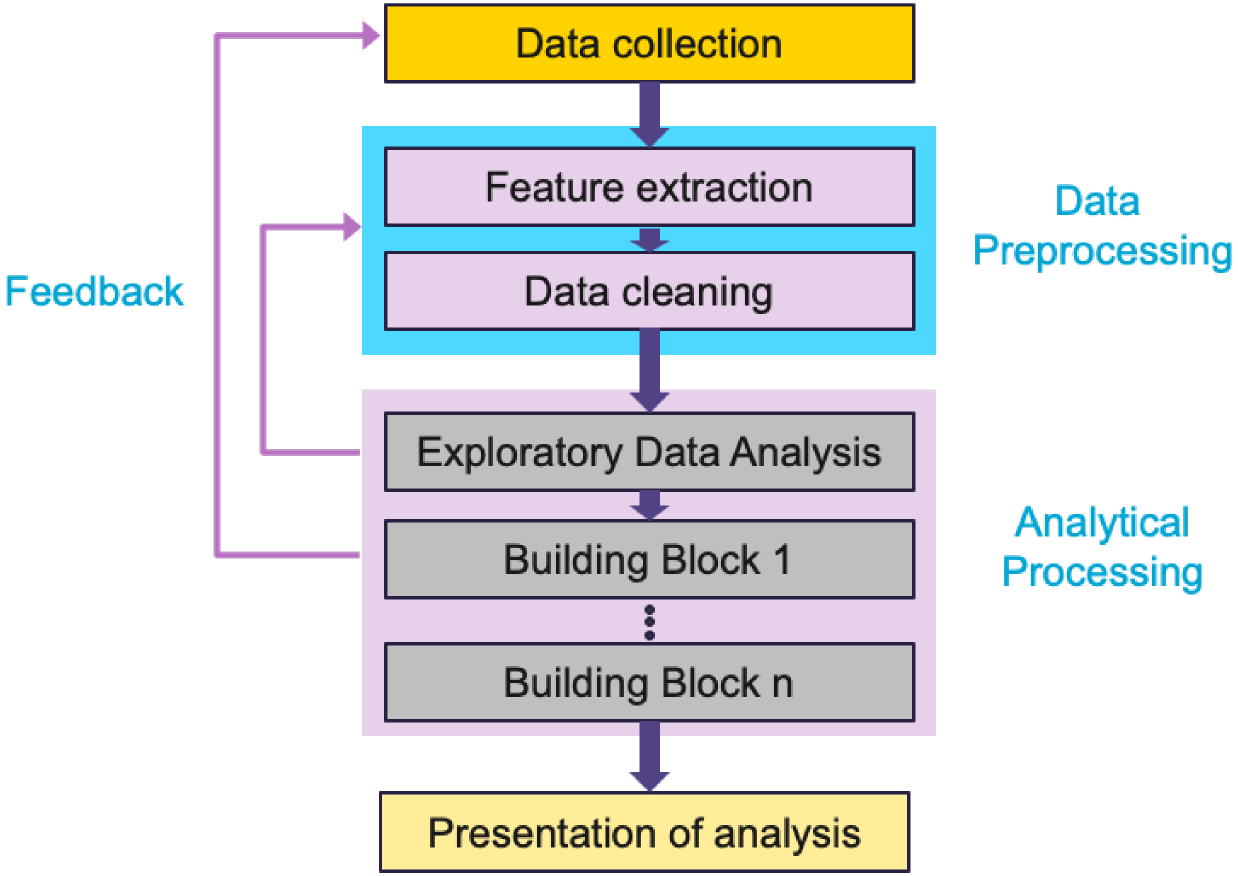
Finding Patterns in Data

Knowledge discovery in databases

The study of collecting, cleaning, processing, analyzing, and gaining useful insights from data

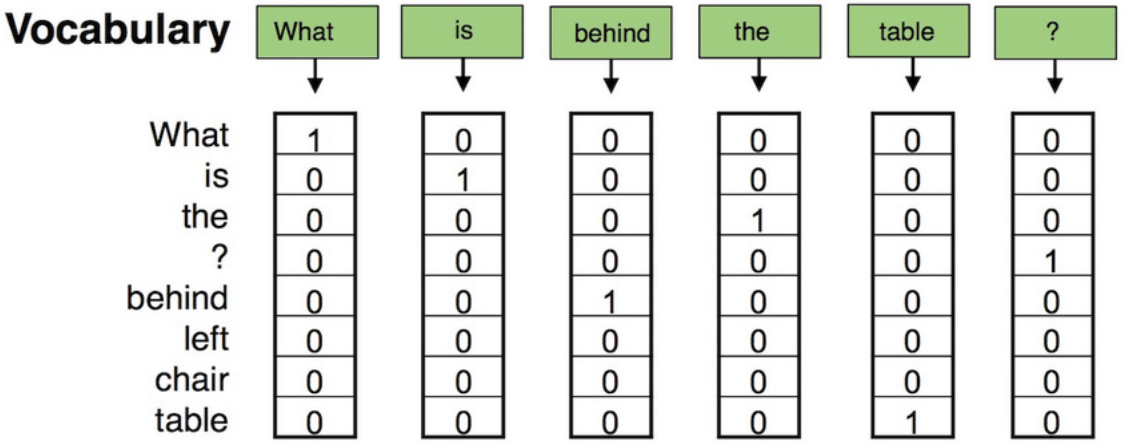
What is Data Mining?

The Data Mining Process

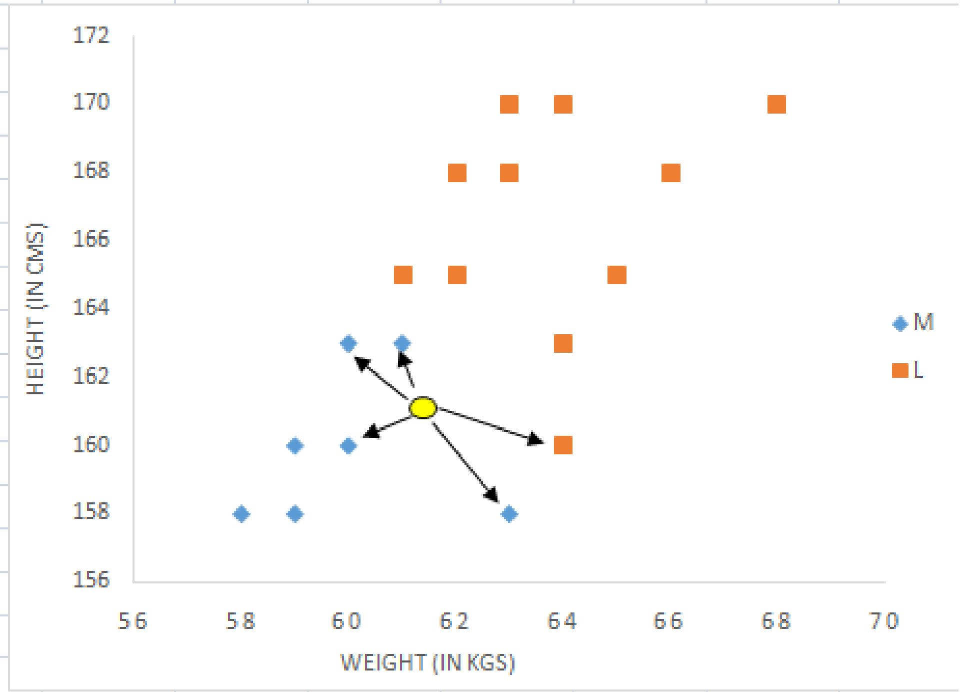


What is Data Mining?

Information Retrieval and Searching by Similarity

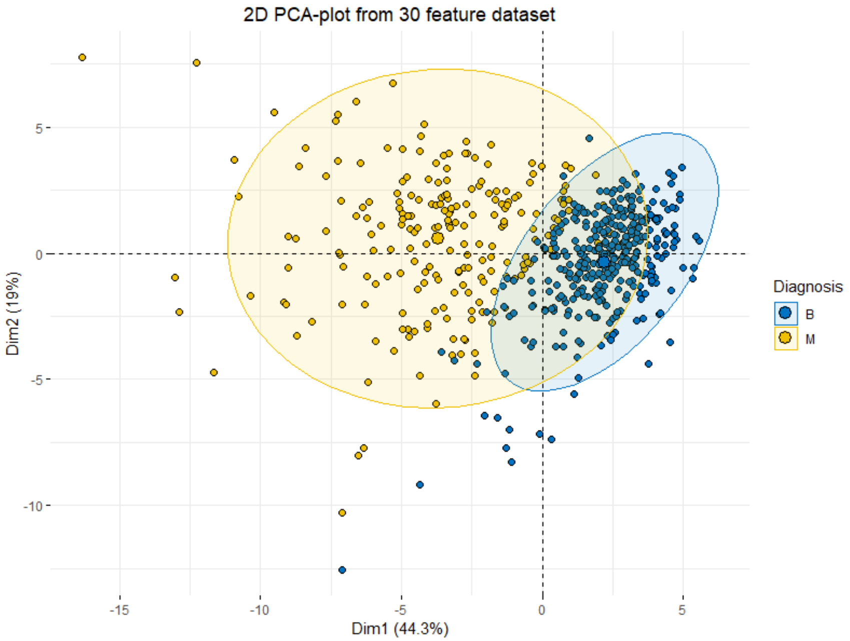


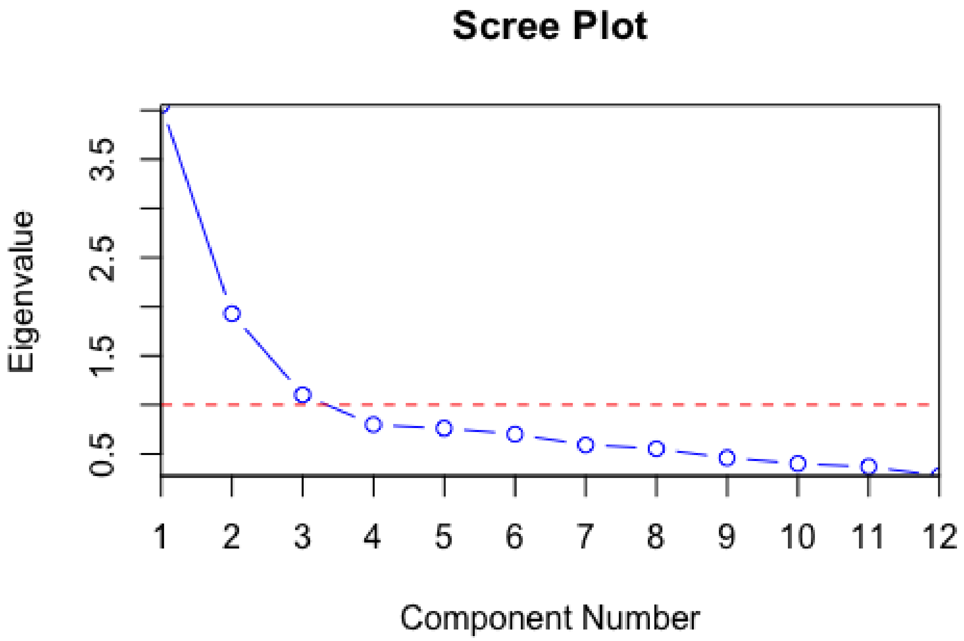
Vocabulary	What	is	behind	the	table	?
What	1	0	0	0	0	0
is	0	1	0	0	0	0
the	0	0	0	1	0	0
?	0	0	0	0	0	1
behind	0	0	1	0	0	0
left	0	0	0	0	0	0
chair	0	0	0	0	0	0
table	0	0	0	0	1	0



What is Data Mining?

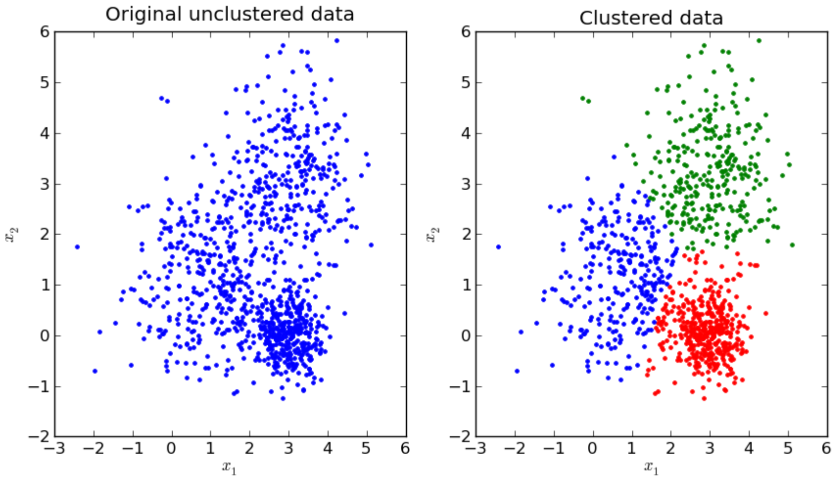
Dimensionality Reduction

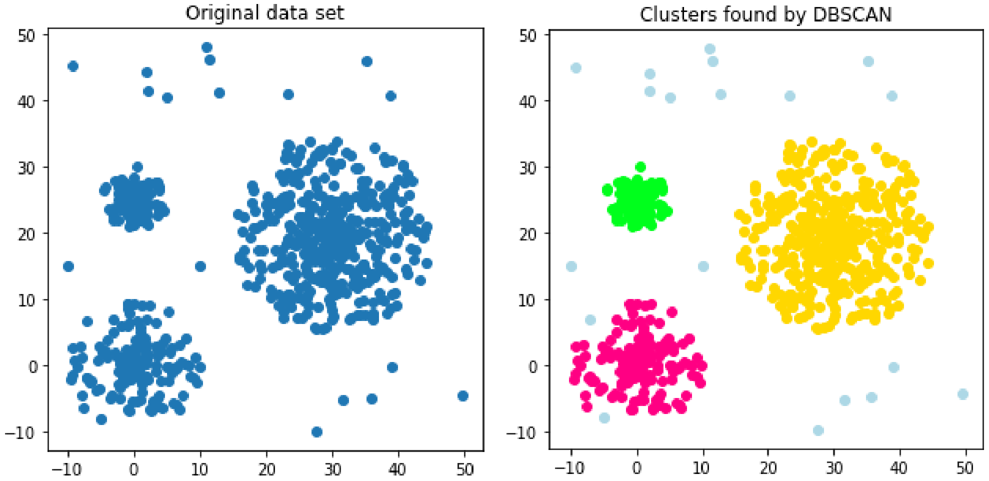

2D PCA-plot from 30 feature dataset


Scree Plot

What is Data Mining?

Clustering



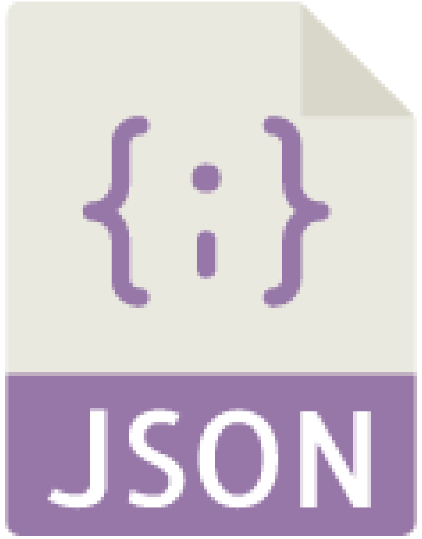


What is Data Wrangling?

Collecting, extracting, transforming, and cleaning data into a form useable for further analysis

What is Data Wrangling?

Working with Different Data Types

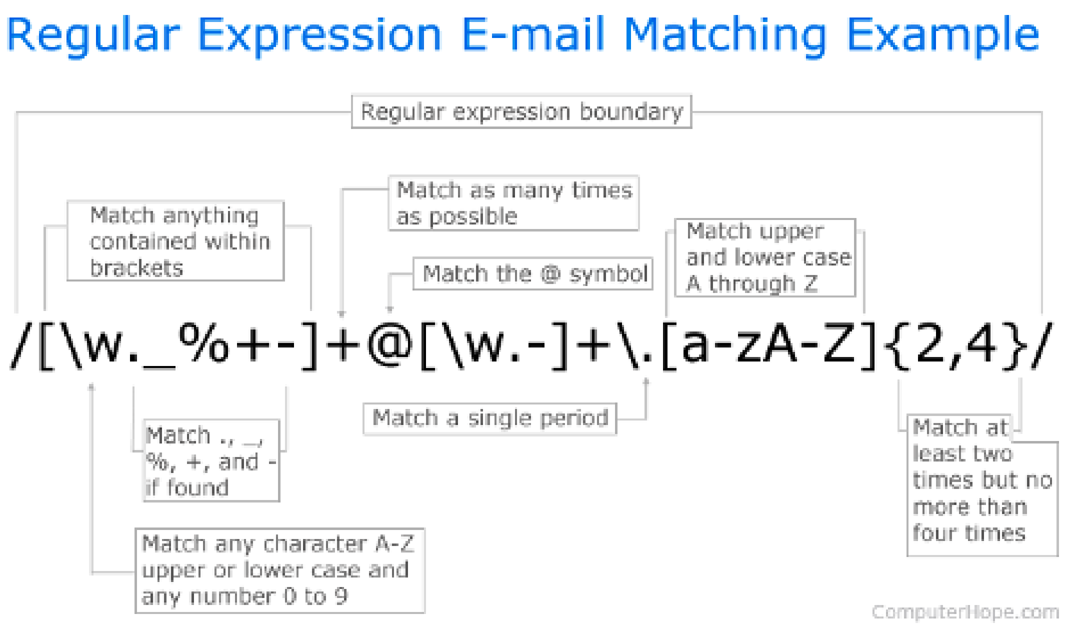


What is Data Wrangling?

Using Regular Expressions

```
/h[a4@](([c<]((k)|(\|<)))|((k)|(\|<))|(x))\s+\
((d)|([t\+]h))[3ea4@]\s+p[l1][a4@]n[3e][t\+]/i
```

(C)2006 FTS Conventures - www.ftsconventures.com



Regular Expression E-mail Matching Example

```
/[\w._%+-]+@[\w.-]+\.[a-zA-Z]{2,4}/
```

Regular expression boundary

Match anything contained within brackets

Match as many times as possible

Match the @ symbol

Match upper and lower case A through Z

Match ., _, %, +, and - if found

Match a single period

Match at least two times but no more than four times

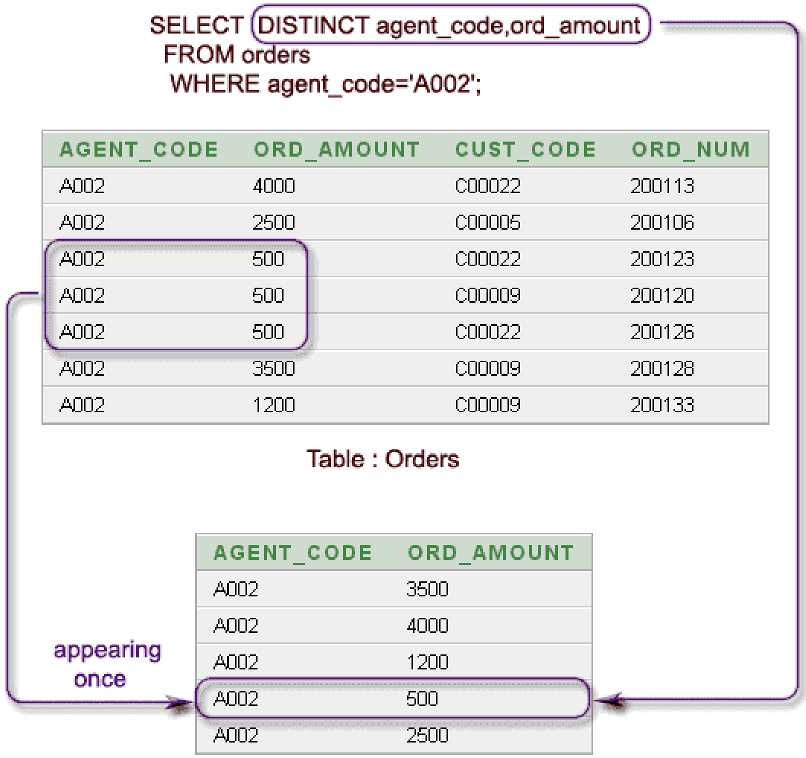
Match any character A-Z upper or lower case and any number 0 to 9

ComputerHope.com

What is Data Wrangling?

Working with Databases

SQL – Structured Query Language



```
SELECT DISTINCT agent_code,ord_amount
 FROM orders
  WHERE agent_code='A002';
```

AGENT_CODE	ORD_AMOUNT	CUST_CODE	ORD_NUM
A002	4000	C00022	200113
A002	2500	C00005	200106
A002	500	C00022	200123
A002	500	C00009	200120
A002	500	C00022	200126
A002	3500	C00009	200128
A002	1200	C00009	200133

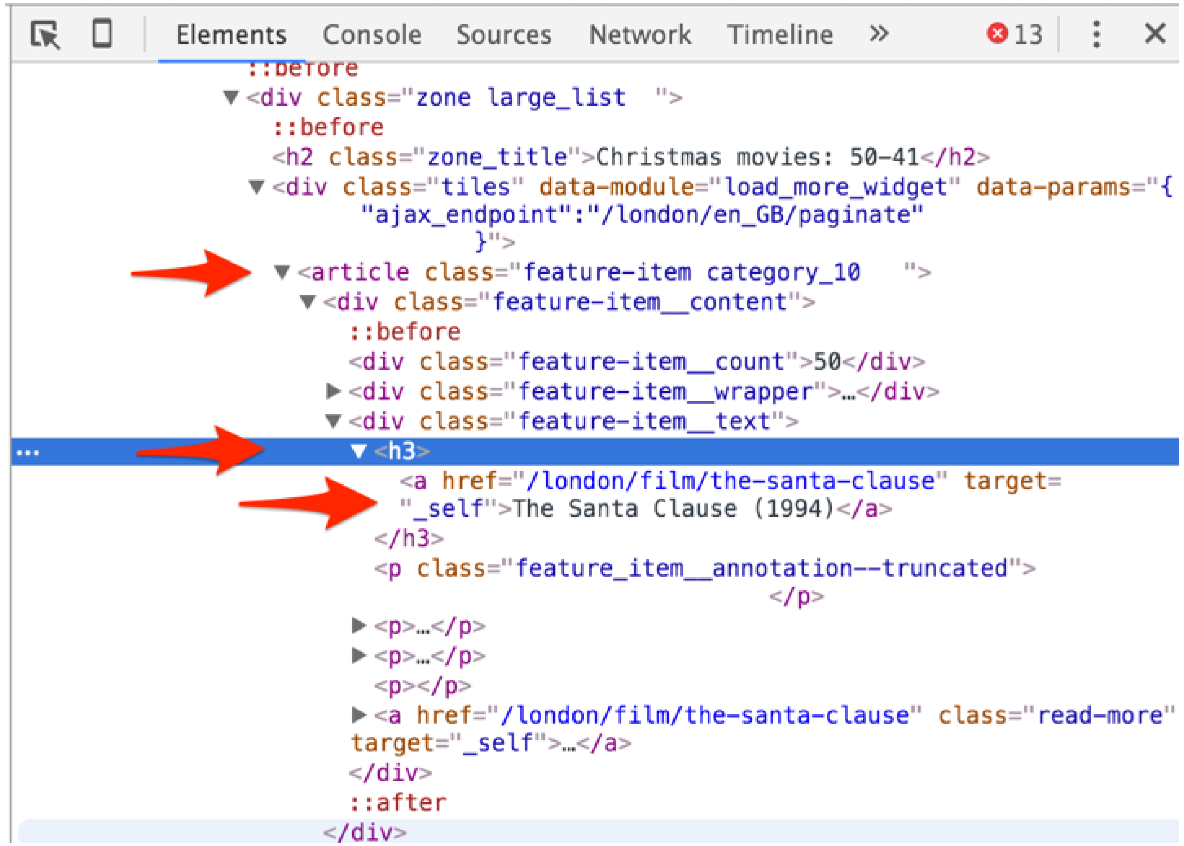
Table : Orders

AGENT_CODE	ORD_AMOUNT
A002	3500
A002	4000
A002	1200
A002	500
A002	2500

Results

What is Data Wrangling?

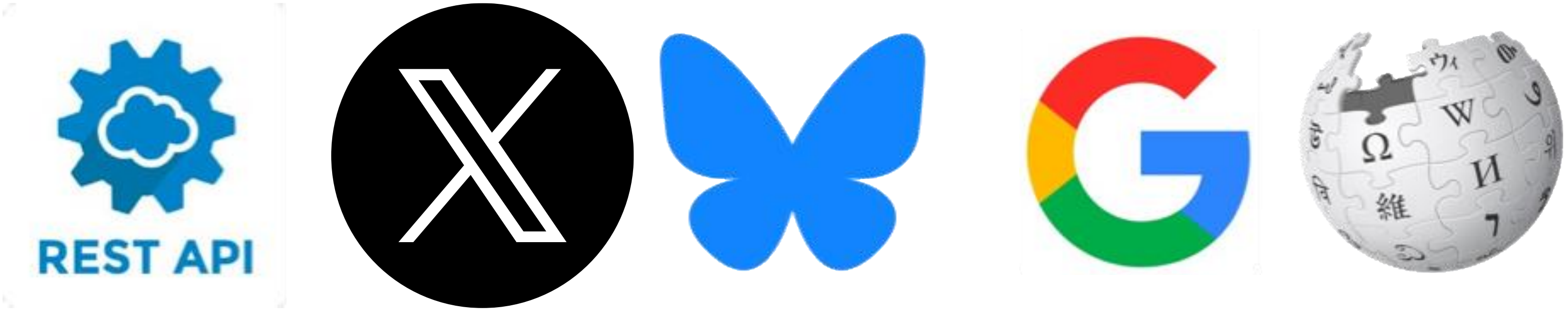
Collecting data through Web Scraping



```
            ::before
          ▼<div class="zone large_list  ">
             ::before
             <h2 class="zone_title">Christmas movies: 50-41</h2>
            ▼<div class="tiles" data-module="load_more_widget" data-params="{
                   "ajax_endpoint":"/london/en_GB/paginate"
                           }">
              ▼<article class="feature-item category_10   ">
                ▼<div class="feature-item__content">
                   ::before
                   <div class="feature-item__count">50</div>
                  ▶<div class="feature-item__wrapper">…</div>
                  ▼<div class="feature-item__text">
…                   ▼<h3>
                       <a href="/london/film/the-santa-clause" target=
                       "_self">The Santa Clause (1994)</a>
                     </h3>
                     <p class="feature_item__annotation--truncated">
                                             </p>
                    ▶<p>…</p>
                    ▶<p>…</p>
                     <p></p>
                    ▶<a href="/london/film/the-santa-clause" class="read-more"
                   target="_self">…</a>
                   </div>
                   ::after
                 </div>
```

What is Data Wrangling?

Collecting data through Web API



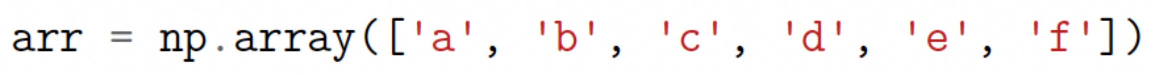
Prologue III
Class Activity 1

AIM

Quick Diagnostics

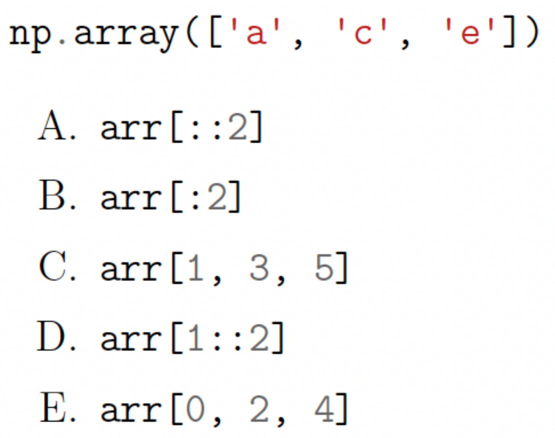
Question 1: Working with Arrays

Given the following array:



```
arr = np.array(['a', 'b', 'c', 'd', 'e', 'f'])
```

Which will yield the following expression?



```
np.array(['a', 'c', 'e'])
```

A. arr[::2]

B. arr[:2]

C. arr[1, 3, 5]

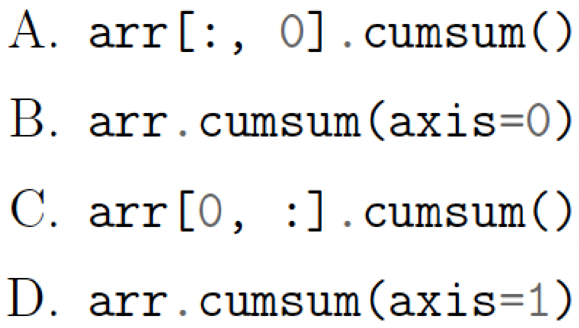
D. arr[1::2]

E. arr[0, 2, 4]

Quick Diagnostics

Question 2: Working with Arrays

Let arr be a 2D ndarray. What is the cumulative sum of arr across rows?



A. `arr[:, 0].cumsum()`

B. `arr.cumsum(axis=0)`

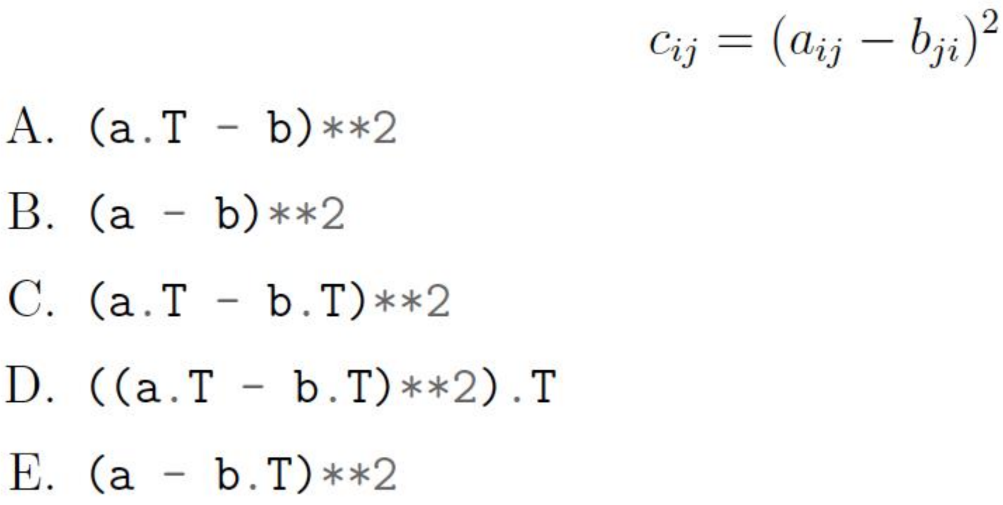
C. `arr[0, :].cumsum()`

D. `arr.cumsum(axis=1)`

Quick Diagnostics

Question 3: Working with Arrays

Let a and b be 2D ndarrays of the same square shape. Which of the following will yield another ndarray c whose elements are given by:



$$c_{ij} = (a_{ij} - b_{ji})^2$$

A. `(a.T - b)**2`

B. `(a - b)**2`

C. `(a.T - b.T)**2`

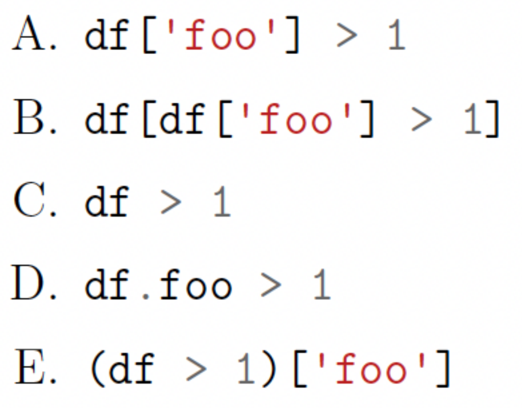
D. `((a.T - b.T)**2).T`

E. `(a - b.T)**2`

Quick Diagnostics

Question 4: Working with pandas (selection)

Let df be a DataFrame. Which of the following expressions will only return rows with value of column foo greater than 1?



A. `df['foo'] > 1`

B. `df[df['foo'] > 1]`

C. `df > 1`

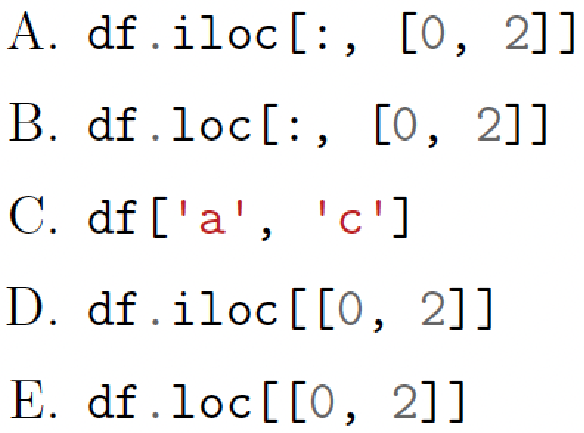
D. `df.foo > 1`

E. `(df > 1)['foo']`

Quick Diagnostics

Question 5: Working with pandas (selection)

Let df be a DataFrame with columns [‘a’, ‘b’, ‘c’, ‘d’, ‘e’]. Which of the following expressions will select only columns ‘a’ and ‘c’?



A. `df.iloc[:, [0, 2]]`

B. `df.loc[:, [0, 2]]`

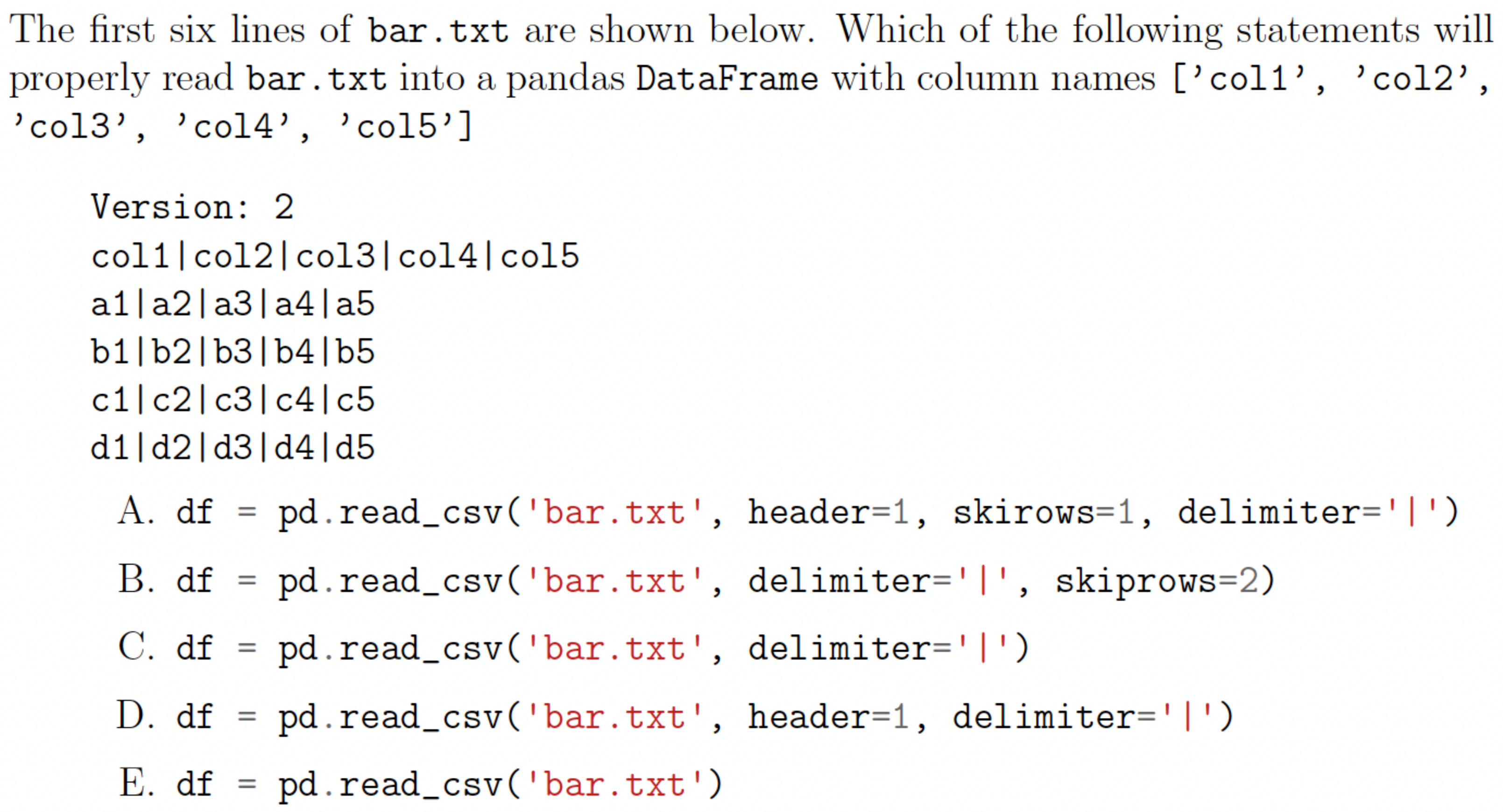
C. `df['a', 'c']`

D. `df.iloc[[0, 2]]`

E. `df.loc[[0, 2]]`

Quick Diagnostics

Question 6: Working with pandas (csvs)



The first six lines of `bar.txt` are shown below. Which of the following statements will properly read `bar.txt` into a pandas `DataFrame` with column names `[’col1’, ’col2’, ’col3’, ’col4’, ’col5’]`

```
Version: 2
col1|col2|col3|col4|col5
a1|a2|a3|a4|a5
b1|b2|b3|b4|b5
c1|c2|c3|c4|c5
d1|d2|d3|d4|d5
```

A. `df = pd.read_csv('bar.txt', header=1, skirows=1, delimiter='|')`

B. `df = pd.read_csv('bar.txt', delimiter='|', skiprows=2)`

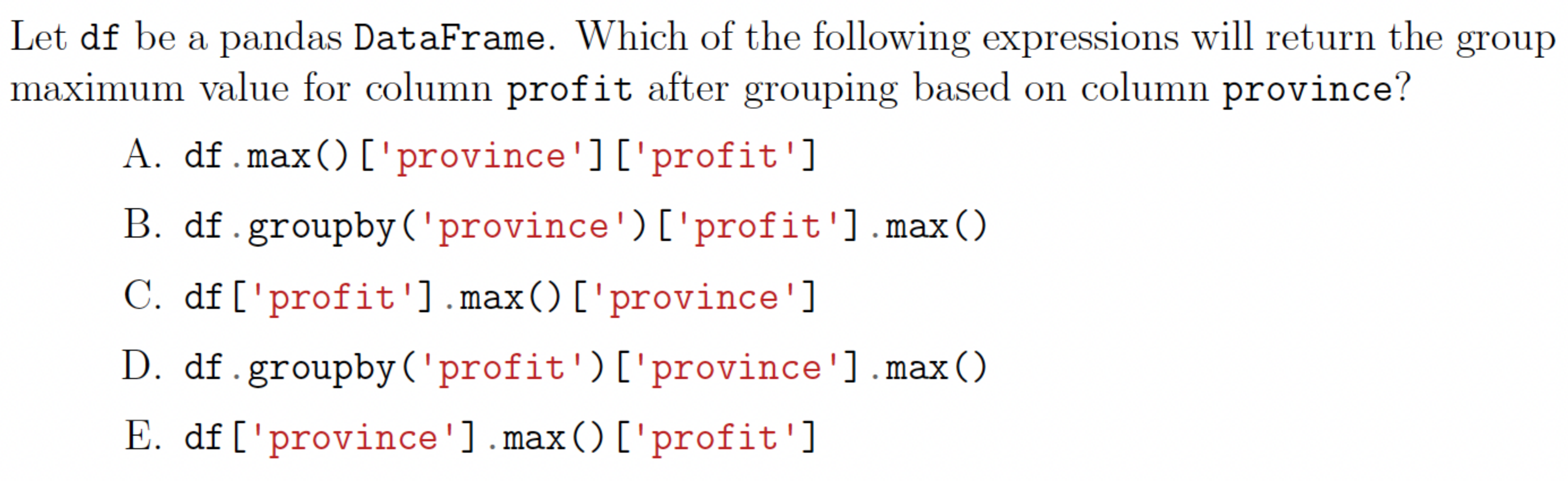
C. `df = pd.read_csv('bar.txt', delimiter='|')`

D. `df = pd.read_csv('bar.txt', header=1, delimiter='|')`

E. `df = pd.read_csv('bar.txt')`

Quick Diagnostics

Question 7: Working with pandas (groupby)



Let `df` be a pandas `DataFrame`. Which of the following expressions will return the group maximum value for column `profit` after grouping based on column `province`?

A. `df.max()['province']['profit']`

B. `df.groupby('province')['profit'].max()`

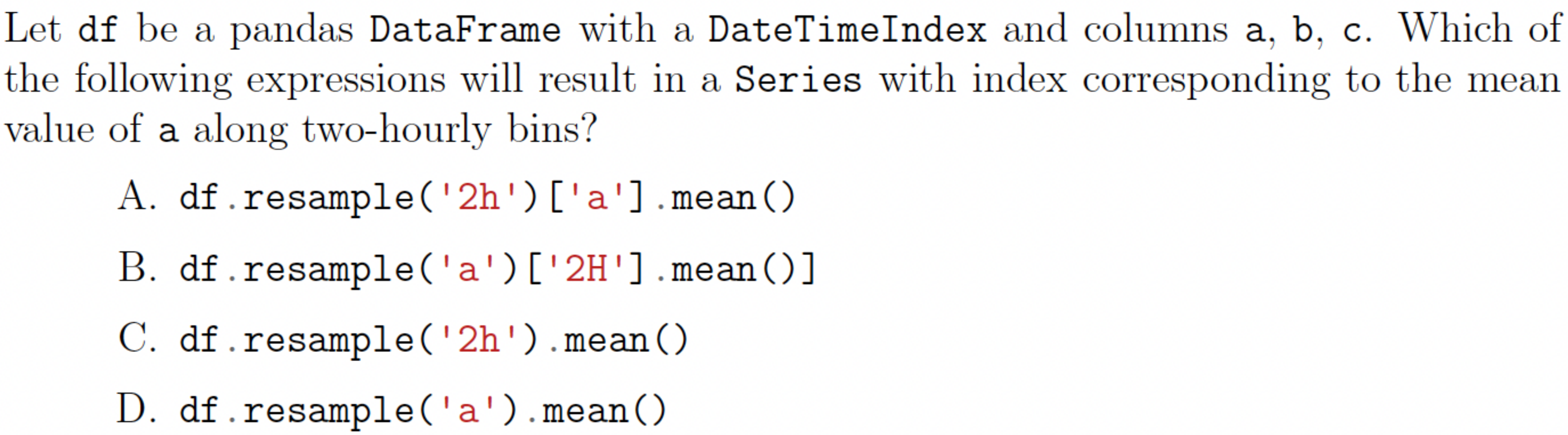
C. `df['profit'].max()['province']`

D. `df.groupby('profit')['province'].max()`

E. `df['province'].max()['profit']`

Quick Diagnostics

Question 8: Working with pandas (timeseries)



Let `df` be a pandas `DataFrame` with a `DateTimeIndex` and columns `a`, `b`, `c`. Which of the following expressions will result in a `Series` with index corresponding to the mean value of `a` along two-hourly bins?

A. `df.resample('2h')['a'].mean()`

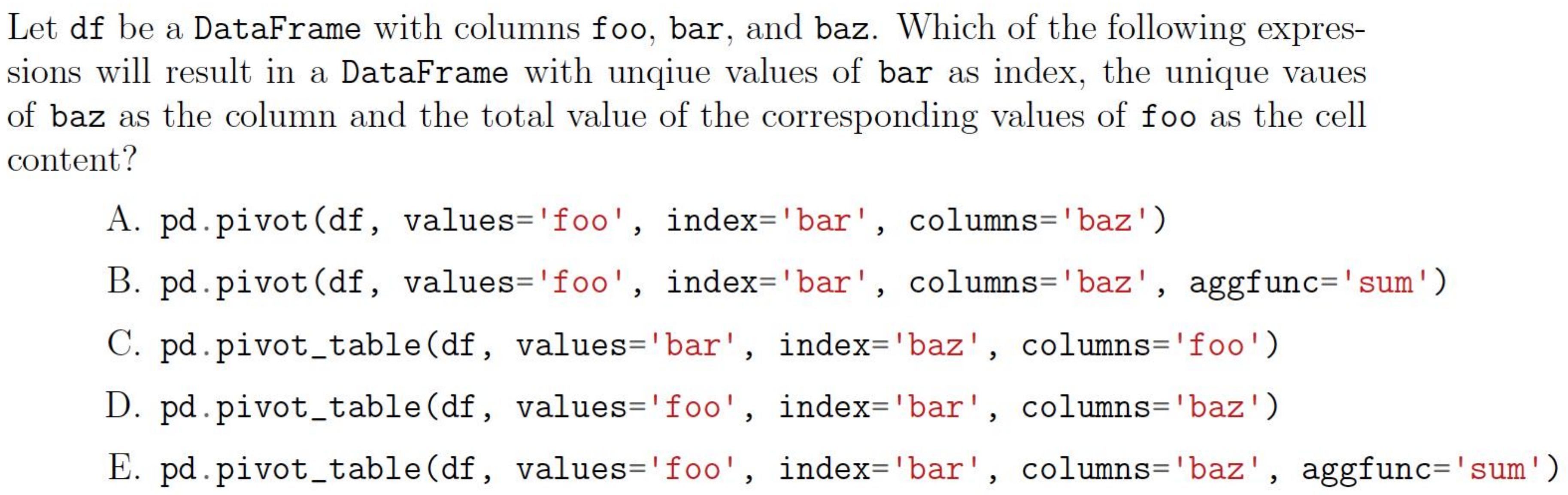
B. `df.resample('a')['2H'].mean()]`

C. `df.resample('2h').mean()`

D. `df.resample('a').mean()`

Quick Diagnostics

Question 9: Working with pandas (pivot)



Let `df` be a `DataFrame` with columns `foo`, `bar`, and `baz`. Which of the following expressions will result in a `DataFrame` with unqiue values of `bar` as index, the unique vaues of `baz` as the column and the total value of the corresponding values of `foo` as the cell content?

A. `pd.pivot(df, values='foo', index='bar', columns='baz')`

B. `pd.pivot(df, values='foo', index='bar', columns='baz', aggfunc='sum')`

C. `pd.pivot_table(df, values='bar', index='baz', columns='foo')`

D. `pd.pivot_table(df, values='foo', index='bar', columns='baz')`

E. `pd.pivot_table(df, values='foo', index='bar', columns='baz', aggfunc='sum')`

Quick Diagnostics

Question 10: Bonus

What is your favorite PDS concept/lesson/library? Why?