1. Select the data format for each file listed on the left:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CSV | XML | JSON | HTML |
| <html>  <head>  <meta http-equiv=”Content-Type” content=”text/html”/>  </head>  <body>  <i>Wolf,4,pack</i>  </body>  </html> |  |  |  |  |
| Wolf,4,pack  Duck,3,flock  Lizard,12,lounge |  |  |  |  |
| <Animal>  <Name>Wolf</Name>  <Number>43</Number>  <Group>pack</DateBorn>  </Animal> |  |  |  |  |
| {  "Name" : "Wolf",  "Number" : 23,  "Group" : “pack"  } |  |  |  |  |

1. What is the result of execution of the following program:

import re

s = 'Abca deaf abck'

re.findall('.[a-f] ',s)

[‘Abca ’, ’deaf ‘]

[‘ca ‘, ‘af ‘]

[‘ca ‘, ‘af ‘, ‘ck ‘]

[‘ca’, ’af’]

[‘ca, ’af’, ‘ck’]

1. We have a retail database which contains two tables as described in the following relational schema:

**Shop**

**Sale**

**ID**

Name (U)

City

State

Zip

IsOpen

**ID**

Item

Price

Date

Transaction (U)

ShopID (FK)

Check statements that correctly describe the above schema:

Each shop should have at least one sale.

Each shop may have many sales

Each sale should be done in only one shop

ShopID is a foreign key in Sale table

Date is a unique key in Sale table

1. Select the correspondence of SQL query and the output

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total number of sales in still opened shops | Total number of sales in a single shop | Income of all shops over 3 months | Total number of sales in all shops |
| SELECT COUNT(Sale.ID)  FROM Sale |  |  |  |  |
| SELECT COUNT(Sale.ID)  FROM Sale, Shop  WHERE Sale.ShopID = Shop.ID  AND Shop.IsOpen = 1 |  |  |  |  |
| SELECT COUNT(Sale.ID)  FROM Sale, Shop  WHERE Sale.ShopID = Shop.ID  AND Shop.Name = “Best shop” |  |  |  |  |
| SELECT SUM(Price)  FROM Sale  WHERE Date BETWEEN ‘2017-01-01’ AND ‘2017-03-31’ |  |  |  |  |

1. What is the result of the execution of the following program:

import numpy as np

a = np.array([1,0,1,0,1,0])

x = a[1:-2]

b = a.reshape(2,3)

y = b[1]

print(x.dot(y))

0

1

[[1,0,1],

[0,1,0]]

[[0,1,0],

[1,0,1]]

[[1,0],

[1,0],

[1,0]]

1. Our data is arranged in Pandas DataFrame (df) in the following way:

|  | **a** | **b** |
| --- | --- | --- |
| **0** | NaN | 1.0 |
| **1** | NaN | NaN |
| **2** | 1.0 | NaN |
| **3** | NaN | 2.0 |

Select the correspondence of the execution of the following commands:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | **a** | **b** | | --- | --- | --- | | **0** | 0.0 | 1.0 | | **1** | 0.0 | 0.0 | | **2** | 1.0 | 0.0 | | **3** | 0.0 | 2.0 | | |  | **a** | **b** | | --- | --- | --- | | 0 NaN  1 NaN  2 1.0  3 NaN  Name: a, dtype: float64 | a 1.0  b 3.0  dtype: float64 |
| df.dropna() |  |  |  |  |
| df.fillna(0) |  |  |  |  |
| df['a'] |  |  |  |  |
| df.apply(lambda x: x.sum()) |  |  |  |  |

1. Select the correspondence between the results of an A/B test and the conclusion:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Launch new technology | Need more tests | Test results are statistically insignificant | Test results are practically insignificant |
| dmin  dmax |  |  |  |  |

1. Where to use each of these techniques:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Find | Collect | Store | Verify | Clean | Analyze |
| Google |  |  |  |  |  |  |
| API |  |  |  |  |  |  |
| Distributions |  |  |  |  |  |  |
| SQL |  |  |  |  |  |  |
| CSV |  |  |  |  |  |  |
| Pandas |  |  |  |  |  |  |
| Manual checking |  |  |  |  |  |  |

1. Choose best plot to represent the following data:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Scatter | Line | Barchart | Histogram | Box plot | Heatmap | Animation | Interactivity |
| Stock changing in time |  |  |  |  |  |  |  |  |
| Correlation between continuous variables |  |  |  |  |  |  |  |  |
| Correlation between categorical variables |  |  |  |  |  |  |  |  |
| Correlation between categorical and continuous variables |  |  |  |  |  |  |  |  |
| 4D data |  |  |  |  |  |  |  |  |
| 5D+ data |  |  |  |  |  |  |  |  |
| Distribution of a single discrete variable |  |  |  |  |  |  |  |  |
| Amount in each category |  |  |  |  |  |  |  |  |