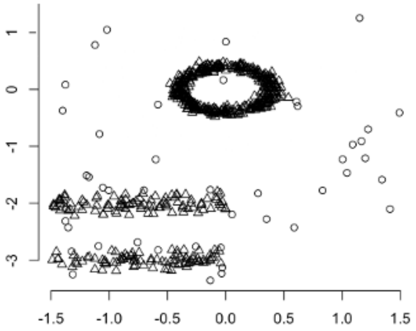
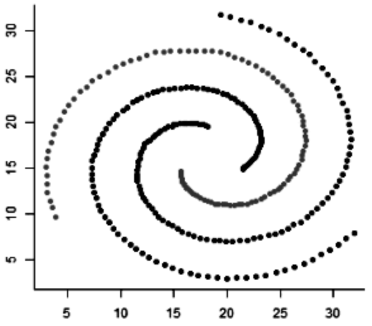
Question 1 (10 marks):

Short-answer questions: answer each of the following questions in a short paragraph.

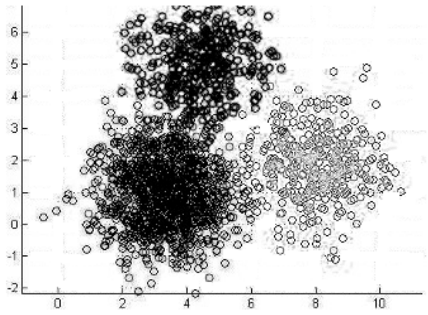
1) Given the following datasets:



(A)



(B)



(C)

I. If we want to apply clustering technique on ***each*** dataset, would it be better to apply ***k-means*** or ***DBSCAN***? And explain why? (4 marks)

Answer:

- (A) DBSCAN is better. (1 mark)
- (B) DBSCAN is better. (1 mark)
- (C) K-means or DBSCAN. (1 mark)
- Because DBSCAN doesn't assume a cluster shape, while Kmeans is suitable for spherical clusters. (1 mark)

II. In figure (A), you can observe some noise in the dataset. (3 marks)

- Which ***step(s)*** in the typical Data Science process will help to identify and fix this noise?
- Briefly explain each step.
- Clearly indicate the order of the ***step(s)*** as part of your answer.

Answer:

- Data preparation / data exploration (1 mark)
- Order: Data preparation / data exploration (1 mark)
- There should be detailed explanation for each step (1 mark)

2) Suppose you have a data set that includes two ***categorical*** and three ***numerical*** columns***.*** (If you don't know the name, you can sketch an example picture.) (3 marks)

i) Name two kinds of graphs that can be used to visualise ***categorical*** data

Answer: Barplot, pie chart (0.5 marks each)

ii) Propose a simple analysis to explore the relationship between a ***categorical*** and a ***numerical*** column.

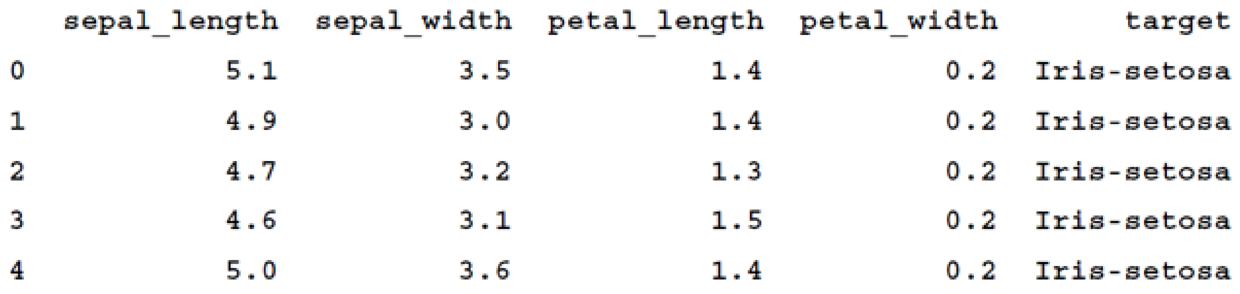
Answer: Boxplot by category (1 mark)

iii) Propose a simple analysis to explore the relationship between two ***numerical*** columns.

Answer: Scatterplot (1 mark)

Question 2 (4 marks):

Considering the following ***iris*** dataset to train a classifier. The attributes are *sepal_length, sepal_width, petal_length and petal_width*. The class labels are in *'target'* column. The datasets contains 150 observations: the first 50 observations are for the type of '***Iris-setosa***', the middle 50 observations are for '***Iris-virginica***', and the last 50 observations are for '***Iris-versicolor***'.



```
   sepal_length  sepal_width  petal_length  petal_width       target
0           5.1          3.5           1.4          0.2  Iris-setosa
1           4.9          3.0           1.4          0.2  Iris-setosa
2           4.7          3.2           1.3          0.2  Iris-setosa
3           4.6          3.1           1.5          0.2  Iris-setosa
4           5.0          3.6           1.4          0.2  Iris-setosa
```

It is required to train a classifier with 3-fold cross validation. Please answer the following questions with plain English, and explain (you may draw diagrams to explain).

1. What are the necessary step(s) to preprocess the data, and explain why preprocessing is important. (2 marks)

   Answer:

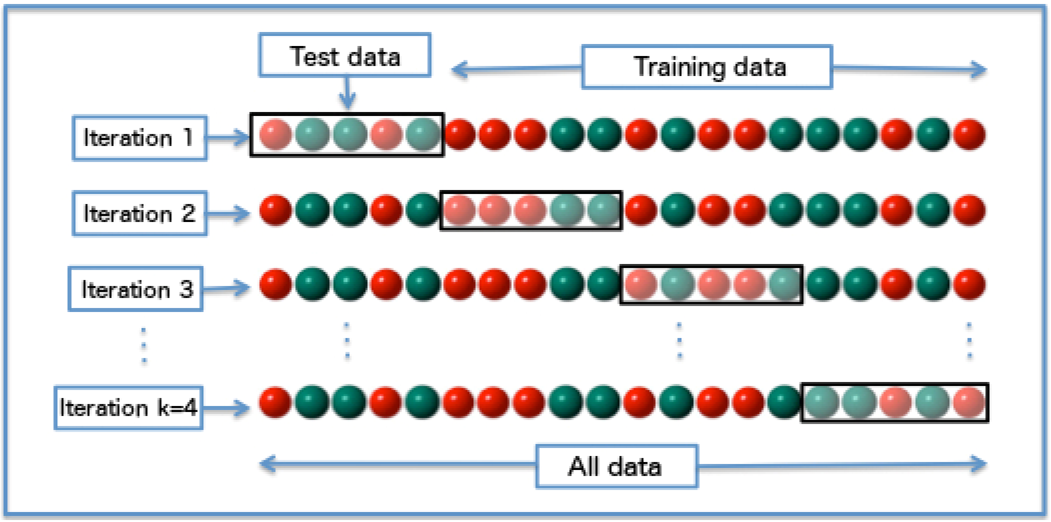
   - Checking errors (outliers or typos) in data entry, which result in few observations having different logic. Extra white spaces, which will affect string comparison. Impossible values, e.g. negative age. (0.5 mark, should mention at least two steps)

- Prepare the data by managing the order of the rows, as they are listed by different classes. If not randomizing them, it will cause the training/test data not reliable. (1 mark)
- Data preprocessing is important because it will help models to perform better keeping in mind that “Garbage in equals garbage out”. (0.5 mark)

2. Apply 3-fold cross validation to the dataset, explaining the process step by step. You may wish to include a diagram as ***part*** of your answer. (2 marks)

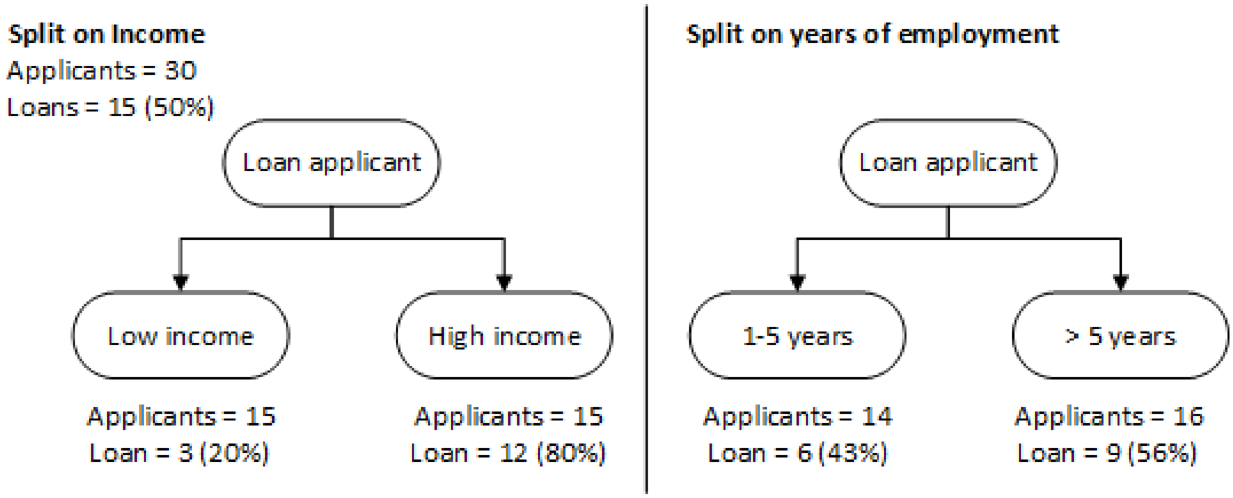
Answer:

- Split the data into 3 equal folds. Use 2 for training and 1 testing, iteratively. (1 mark)
- K-fold cross validation is useful with small datasets. (1 mark)



Question 3 (3 marks):

Consider we have a sample of 30 loan applicants with two variables *Income range (Low/ High)* and *Years of employment(1-5/ >5)*. 15 out of these 30 were granted the loan. Now, we want to build a ***Decision Tree*** on this data. In the figure below, we split the population using the two input variables *Income* and *Years of employment*.



Which split is producing more homogeneous sub-nodes using the *Gini* index (equation is given at the end of the exam paper)? and explain why. (3 marks)

Answer:

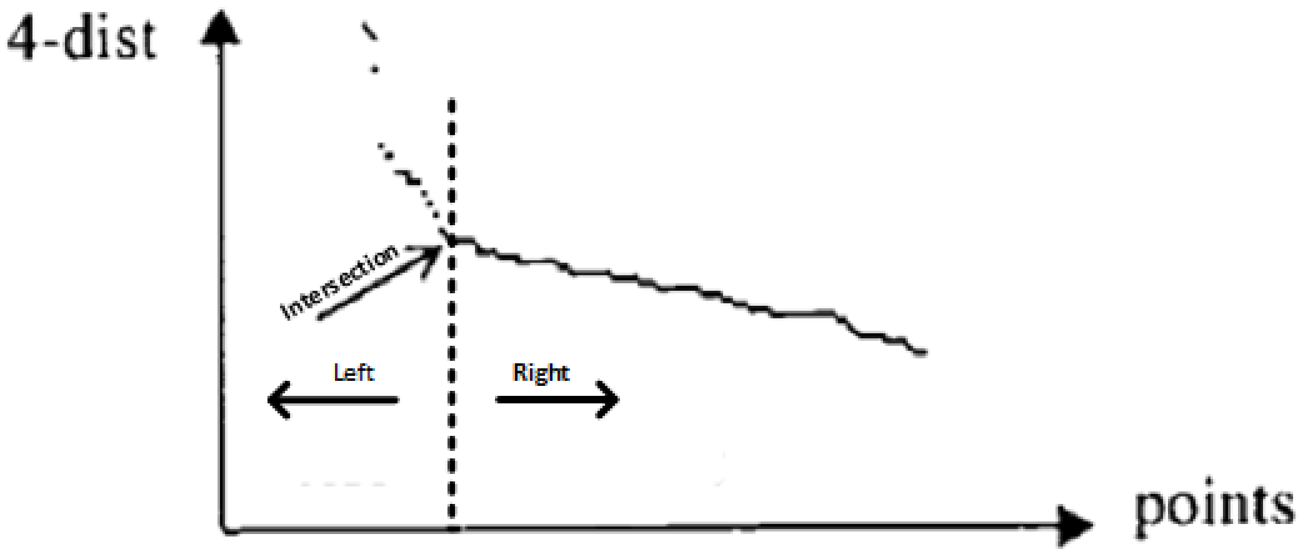
- For split on Income: (1 mark)
  Gini for sub-node Low income = $0.2^2 + 0.8^2$ = 0.68
  Gini for sub-node High income = $0.8^2 + 0.2^2$ = 0.68
  Calculate weighted Gini for Split Income = (15/30) * 0.68 + (15/30) * 0.68 = 0.68

- For split on Years of employment: (1 mark)
  Gini for sub-node 1-5 years = $0.43^2 + 0.57^2$ = 0.51
  Gini for sub-node >5 years = $0.56^2 + 0.44^2$ = 0.51
  Calculate weighted Gini for Split Employment = (14/30)*0.51+(16/30)*0.51 = 0.51

- Splitting on Income is better as its Gini index is higher. (1 mark)

Question 4 (4 marks):

The following figure shows the *k* ***distance graph*** for a DBSCAN algorithm where *minPts* is equal to 4:



Please answer

1. What is the meaning of the ***intersection*** point in the graph. How can we use it in the DBSCAN algorithm? (2 marks)

   Answer:

   - The threshold point represents the value of the Eps parameter. Eps is the maximum distance required to consider a neighbour point. (1 mark)

   - Good values of Eps are where this plot shows a strong bend. (1 mark)

2. What is the meaning of the points to the ***Left*** of the dotted line in this graph? (1 mark)

   Answer: Noise. If Eps is chosen less than the noise value, a large part of the data will not be clustered and will be considered noise

3. What is the meaning of the points to the ***Right*** of the dotted line in this graph? (1 mark)

   Answer: Clusters. If Eps is chosen greater than the noise value, the model will detect clusters. However, a too high value of Eps will merge and the majority of objects will be in the same cluster.

Question 5 (5 marks):

Consider the following situation: In a local government's road traffic data, the severity of traffic accidents is defined by how many people are injured in the accident. Consider training a classifier on this data, specifically, the following columns are included as the data source: *time, day in a week, date, driver's age, the number of people injured, location, severity*. The target class is *'severity'*.

Please answer

1. In this data source, ***identify*** and ***explain*** which column has the potential to cause data leakage? (3 marks)

   Answer:
   - The column ‘the number of people injured’ can cause potential data leakage, (1.5 marks)
   - because the target class ‘severity’ is defined by how many people are injured in the accident. (1.5 marks)

2. What is the impact of data leakage on the trained classifier? (1 mark)

   Answer: Data Leakage creates unexpected additional information in the training data, allowing a model or machine learning algorithm to make unrealistically good predictions. It is also known as overfitting.

3. How to avoid data leakage while training the severity of traffic accidents classifier? (1 mark)

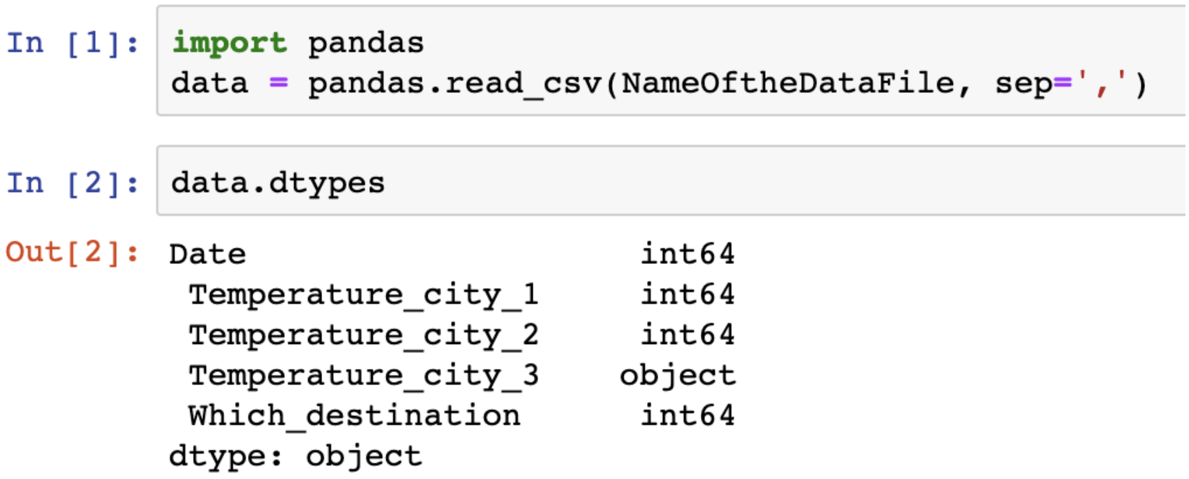
   Answer: Exclude the column ‘the number of people injured’ from training.

Question 6: (7 marks)

The following table contains records of a travel agency. According to the temperature of three popular destinations, they record whether the user picks the first, second, or the third destination. The following figure shows the records they have in their source data. Temperature is in Fahrenheit.

```
Date, Temperature_city_1, Temperature_city_1, Temperature_city_1, Which_destination
20140910,  80,  32, 40, 1
20140911, 100,  50, 36, 2
20140912, 102,  55, , 1
20140912,  60,  20, 35, 3
20140914,  60,  30, 32, 3
20140914, 800,  57, 42, 2
```

After running the following python code:



```
In [1]:  import pandas
         data = pandas.read_csv(NameOftheDataFile, sep=',')

In [2]:  data.dtypes

Out[2]:  Date                      int64
          Temperature_city_1      int64
          Temperature_city_2      int64
          Temperature_city_3     object
          Which_destination       int64
         dtype: object
```

Check the above data, Python code and output carefully, and answer the following questions:

- What error(s) you observed from this given source data? (2 mark)

  Answer: There are two outliers/noise values: one in row 3 for the 3nd city (missing value), and the other is in row 6 for the 1st city (temperature can't go up to 800 degrees) (1 mark each)

- Given that we do not want to ignore the error-affected rows (observations) and do not want to fix them by filling a constant value (or null), please list at least two methods to fix the above identified error(s), and compare which one is better? (You do not need to answer this via Python code, and just answer in plain English) (2 marks)

  Answer:
  - One method is to use the mean/median value of the corresponding column, and the other is to use the mean/median value of the corresponding row (values across the three city columns only). (1 mark)

  - Mean/median value of the column is better because it's based on the temperatures of the same city. For instance, city 1 tend to have higher temperatures than the other two. (1 mark)

- After loading the data with *pandas.read_csv()*, what error(s) you observed from the loaded data? Then explain how this will affect the future data analysis and modelling? (you might use examples to explain) (3 marks)

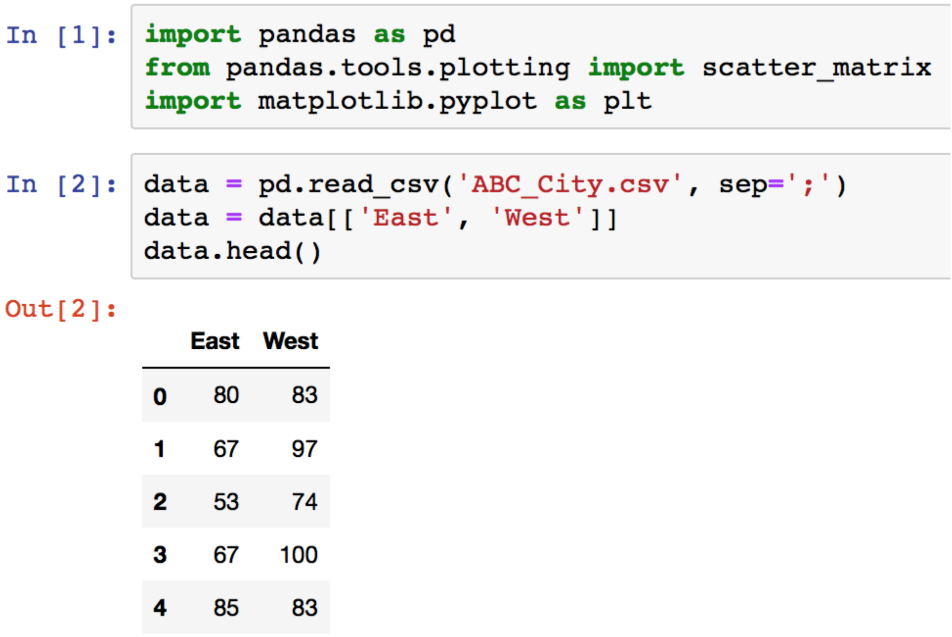
  Answer:
  - The first column's type was not correctly recognised. It is Int64, but it should be a 'date'. (0.5 mark)

- The 4th column's type was not correctly recognised. It's object, but it should be int64 (0.5 marks)

- This will affect the following analysis and modelling when calculating the distances between rows (observations), then lead to problematic analysis and modelling. (1 mark for each of the above point)

Question 7 (7 marks):

We are investigating the Air Quality Index (AQI) in a specific month between two regions (East and West) in a City named ABC. The AQI is a derived value based on multiple data readings from different pollutants that have different underlying units of measure. Usually, an AQI score above 100 indicates a 'poor' air quality level.



```
In [1]: import pandas as pd
        from pandas.tools.plotting import scatter_matrix
        import matplotlib.pyplot as plt
```

```
In [2]: data = pd.read_csv('ABC_City.csv', sep=';')
        data = data[['East', 'West']]
        data.head()
```

Out[2]:

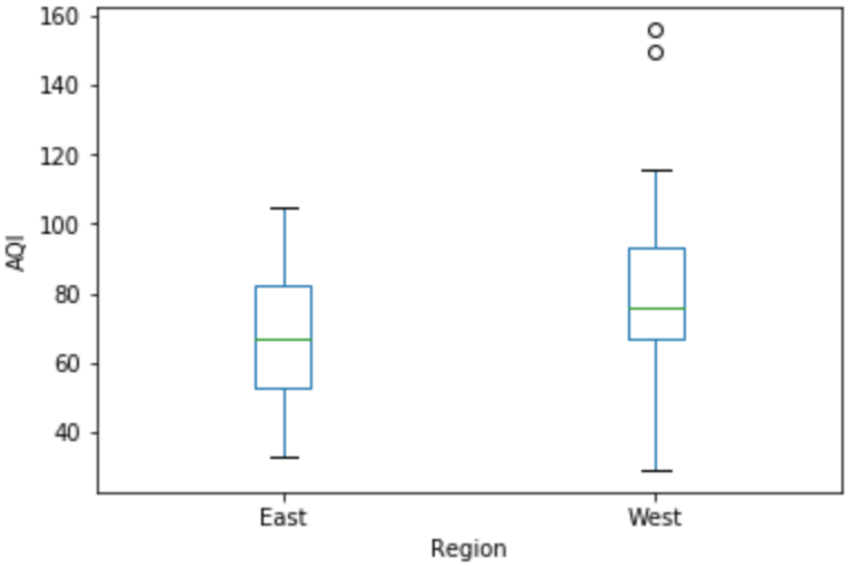
	East	West
0	80	83
1	67	97
2	53	74
3	67	100
4	85	83

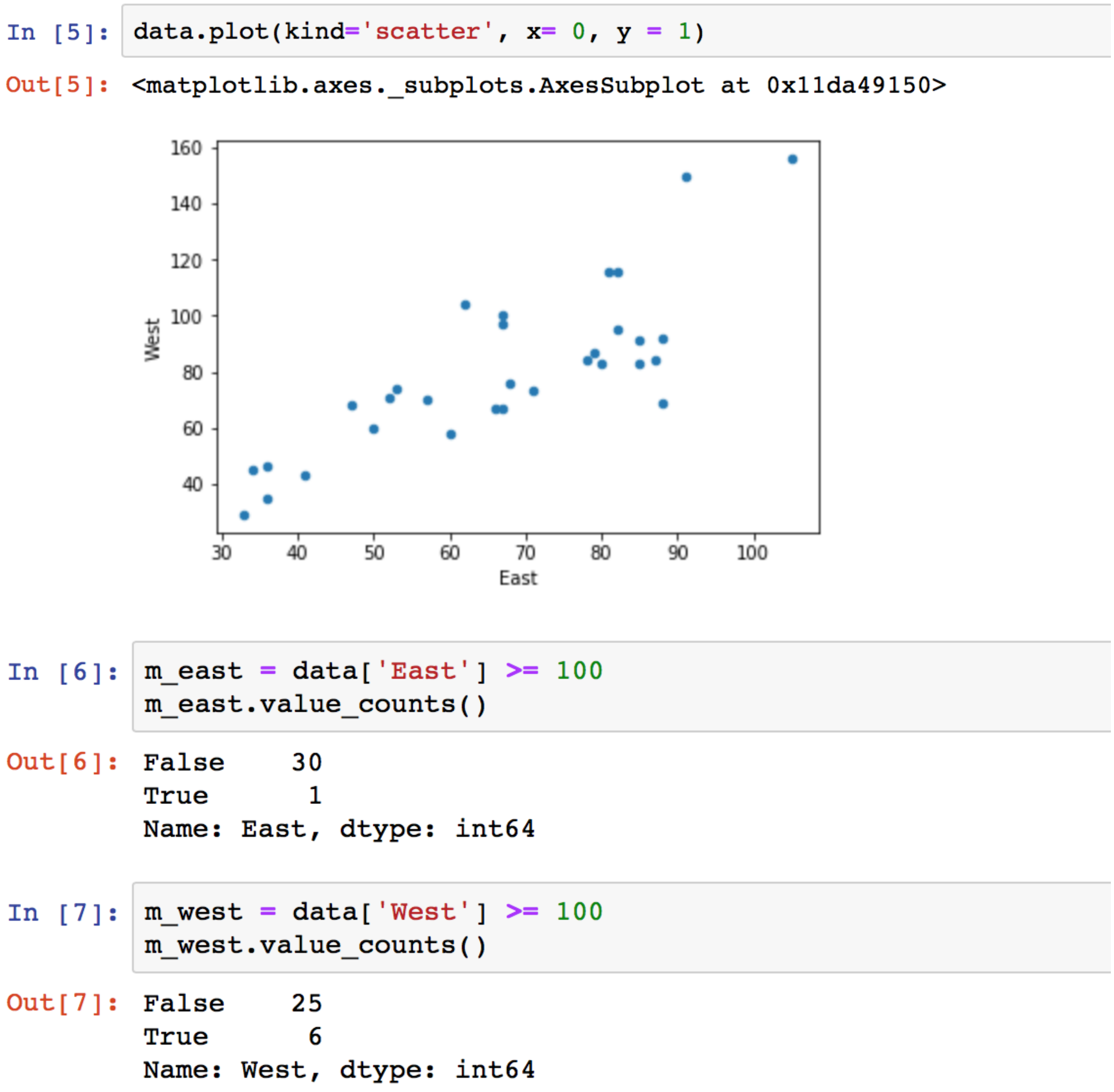
```
In [3]: data.describe()
```

Out[3]:

	East	West
count	31.000000	31.000000
mean	67.032258	80.290323
std	19.300231	29.019756
min	33.000000	29.000000
25%	52.500000	67.000000
50%	67.000000	76.000000
75%	82.000000	93.500000
max	105.000000	156.000000

```
In [4]: data.boxplot(column=['East', 'West'])
        plt.xlabel('Region')
        plt.ylabel('AQI')
        plt.grid(False)
```





```
In [5]: data.plot(kind='scatter', x= 0, y = 1)

Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x11da49150>
```

```
In [6]: m_east = data['East'] >= 100
        m_east.value_counts()

Out[6]: False    30
        True      1
        Name: East, dtype: int64

In [7]: m_west = data['West'] >= 100
        m_west.value_counts()

Out[7]: False    25
        True      6
        Name: West, dtype: int64
```

Check the above Python code and the output, and answer the following questions:

1. What is the mean and the median AQI score for West region of ABC City? (1 mark)

   Answer: Mean: 80.290323; median 76 (0.5 marks each)

2. Comment on the shape of the boxplot for the West region of the ABC City. (1 marks)

   Answer:Positively skewed with 2 outliers. (0.5 marks for skewness, and 0.5 marks for outliers)

3. Answer and explain which of the following value would be better to report as the measure of the West region of ABC City: (1 mark)
    a. 1) mean;
    b. 2) median.

    Answer: Median would be better as it is positively skewed. (1 mark)

4. How many days are the air quality considered ‘poor’ in West and East regions, separately? (1 mark)

    Answer: East: 1; West: 6. (0.5 marks each)

5. Explain the what you can obtain from this scatter plot in terms of the AQI relationship between the West and East regions. (1 mark)

    Answer: Linear positive correlation. (1 mark)

6. Suppose the air quality monitoring instrument was not working on the 15th of this month in the East region, but the AQI value in the West region was recorded as 60. Explain how would you fill this missing value for the East region. (1 mark)

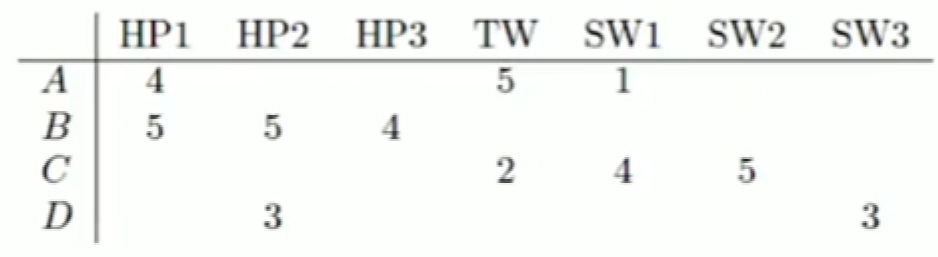
    Answer: Considering the linear positive correlation, we can fill the missing value by using this, (e.g. the value would be around 50). (1 mark)

7. As explained in the lecture, there are six steps in a data science project. Which step(s) should the above Python code belong to? (1 mark)

    Answer: Data Retrieval, Data curation, Data Exploration. (1 mark)

Question 8 (10 marks):

Given the following user-item rating matrix for Collaborative Filtering:



	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Answer the following questions:

1. Calculate the *Jaccard* (see appendix for equations) Similarity: (2 marks)
   a) between user *A* and *B*;
   b) between user *A* and *C*;

2. Calculate the *Cosine* (see appendix for equations) Similarity: (2 marks)
   a) between user *A* and *B*;
   b) between user *A* and *C*;

3. Calculate the *Pearson's correlation coefficient* (see appendix for equations): (2 marks)
   a) between user *A* and *B*;
   b) between user *A* and *C*;

4. Compare the results of the above three similarities, and identify and explain which one is better to measure the closeness between *A* and *B* when considering the missing values. (2 marks)

5. Is *kNN*-based (*k Nearest Neighbour*) Collaborative Filtering technique a personalised recommendation technique or Non-personalised technique, and explain why. (2 marks)

Answers:

1. *sim(A,B) = 0.2; sim(A,C) = 0.5 (1 mark each)*

2. *sim(A,B) = 0.38, sim(A,C) = 0.32 (1 mark each)*

3. *sim(A,B) = 0.09 and sim(A, C) = -0.56 (1 mark each)*

4. *PCC* is better as it treats missing ratings as "average”, and can handle "tough raters" and "easy raters“. Cosine similarity treats missing ratings as negative, while jaccard ignores ratings. *(1 mark for identifying PCC, 1 mark for explanation)*

5. *kNN-based CF is* a personalised recommendation technique. The reason is it can do recommendation based on each users’ rating history. (1 mark for correct answer, and 1 mark for explanation)

********************************END OF EXAM QUESTIONS********************************

Appendix:

1. *Gini* Index: is the sum of square of probability for success and failure (p^2+q^2).

2. *Jaccard* similarity:
   The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

   (If A and B are both empty, we define J(A,B) = 1.)

3. *Cosine* Similarity:
   Given two vectors of attributes, **A** and **B**, the cosine similarity, ***cos(θ)***, is represented using a dot product and magnitude as:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

   where A_i and B_i are components of vector **A** and **B** respectively.

4. *Pearson's correlation coefficient*: Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.
   Given paired data {$(x_1, y_1), \ldots, (x_n, y_n)$} consisting of n pairs, the *Pearson's correlation coefficient* (denoted as r) between x and y is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

   where

   n is sample size,

   x_i, y_i are the individual sample points indexed with i.

   $\bar{x} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.