Question 3 · Data Science

(a) In your own words, explain the benefits of vectorised style of programming for data science

Vectorisation is used to speed up code without using loop. Using such a function can help in maximizing the running time of code efficiency. Vectorisation uses functions that has been written in C. Advantages:

- more concise and easier to read
- fewer lines of code
- closer to mathematical notation.

(b) Rewrite the following code in a vectorised style

```
xs = c(4,1,6,2,9,10)
y=0
for (x in xs){
  if (x%%2==0){
    y = y+x
  }
}
```

```
sum(xs[xs%%2==0])
```

(c) In your own words, what is the main difference between a Scikit-Learn classifier and a regressor?

The score metrics used in each case are the main way to distinguish a classifier from a regressor in scikit-learn. Classifiers would be using metrics such as accuracy, confusion matrix, etc while regressor will be using metrics such as RMSE.

(d) The following data is not tidy. Explain why not, and show what it would look like in tidy data.

Country	Metric	2019	2020
Ireland	Population	5.1	5.2
	GDP	101	102
France	Population	71	72
	GDP	400	420

Country	Metric	Year	Value
Ireland	Population	2019	5.1
Ireland	Population	2020	5.2
Ireland	GDP	2019	101
Ireland	GDP	2020	102
France	Pop	2019	

In tidy data format we should have an observation per row, in this case we have a total of 6 observations in 2-6 rows.

(d) Suppose we have two dplyr tibbles named rentals and customers, as shown below. Notice that not every CustomerID has an entry in the customers table. Write dplyr join to create a tibble containing all rentals together with the corresponding name and addresses. Names and addresses should be blank wherever they are not available.

RENTALS TABLE

Date	Movie ID	Customer ID
01-Jan	102	1
02-Jan	101	2
02-Jan	102	3
05-Jan	103	1
05-Jan	104	7

CUSTOMER TABLE

Customer ID	Name	Address
1	Bob	11, Haight St
2	Frida	Oxford Circus
3	Carrie	99, Fifth Ave

~~dplyr left_join(RENTALS_TABLE, CUSTOMER_TABLE, by = "customer ID")~~

dplyr: left_join(RENTALS TABLE, CUSTOMER TABLE, by = "CUSTOMER ID")