Q1

- You have been hired by the ESA to build a model that predicts the amount of oxygen that an astronaut consumes when performing five minutes of intense physical work. The descriptive features for the model will be the age of the astronaut and their heart rate throughout the work.

  The regression model is:

  ```
  OxyCon = w[0] + w[1] x Age + w[2] x HeartRate
  ```

ID	OxyCon	Age	HeartRate
1	37.99	41	138
2	47.34	42	153
3	44.38	37	151
4	28.17	46	133
5	27.07	48	126
6	37.85	44	145
7	44.72	43	158
8	36.42	46	143
9	31.21	37	138
10	54.85	38	158
11	39.84	43	143
12	30.83	43	138

Q1(a)

Assuming the current weights in a multivariate linear regression model are w[0] = -59.50, w[1] = -0.15 and w[2] = 0.60, make a prediction for each training instance using this model.

```
Prediction(1) = -59.5 - 0.15 x 41 + 0.60 x 138 = 17.15
```

…

ID	OxyCon	Age	HeartRate	Prediction
1	37.99	41	138	17.15
2	47.34	42	153	26.00
3	44.38	37	151	25.55
4	28.17	46	133	13.40
5	27.07	48	126	8.9
6	37.85	44	145	20.90
7	44.72	43	158	28.85
8	36.42	46	143	19.40
9	31.21	37	138	17.75
10	54.85	38	158	29.60
11	39.84	43	143	19.85
12	30.83	43	138	16.85

Q1(b)

Calculate the sum of squared errors (i.e. loss) L_2

ID	OxyCon	Prediction	Error	SSE
1	37.99	17.15	20.84	434.31
2	47.34	26.00	21.34	455.4
3	44.38	25.55	18.83	354.57
4	28.17	13.40	14.77	218.15
5	27.07	8.9	18.17	330.15
6	37.85	20.90	16.95	287.3
7	44.72	28.85	15.87	251.86
8	36.42	19.40	17.02	289.68
9	31.21	17.75	13.46	181.17
10	54.85	29.60	25.25	637.56
11	39.84	19.85	19.99	399.6
12	30.83	16.85	13.98	195.44
			Sum	**4035.19**
			L_2	2017.595

Q1(c)

Assuming a learning rate of .000002, calculate the weights at the next iteration of the gradient descent algorithm and the loss at this point in the algorithm. Is it better or worse?

Weights are adjusted using the following:

$$\mathbf{w}[j] \leftarrow \mathbf{w}[j] + \alpha \sum_{i=1}^{n} ((\overbrace{t_i - \mathbb{M}_{\mathbf{w}}(\mathbf{d}_i)}^{\text{error}}) \times \mathbf{d}_i[j])$$

Q1(c)

Calculate the error delta for each descriptive feature and sum over all features:

ID	OxyCon	Prediction	Age	HeartRate	Error	d[0] x Error	Age x Error	HR x Error
1	37.99	17.15	41	138	20.84	20.84	854.44	2875.92
2	47.34	26.00	42	153	21.34	21.34	896.28	3265.02
3	44.38	25.55	37	151	18.83	18.83	696.71	2843.33
4	28.17	13.40	46	133	14.77	14.77	679.42	1964.41
5	27.07	8.9	48	126	18.17	18.17	872.16	2289.42
6	37.85	20.90	44	145	16.95	16.95	745.8	2457.75
7	44.72	28.85	43	158	15.87	15.87	682.41	2507.46
8	36.42	19.40	46	143	17.02	17.02	782.92	2433.86
9	31.21	17.75	37	138	13.46	13.46	498.02	1857.48
10	54.85	29.60	38	158	25.25	25.25	959.5	3989.5
11	39.84	19.85	43	143	19.99	19.99	859.57	2858.57
12	30.83	16.85	43	138	13.98	13.98	601.14	1929.24
					Sum	**216.47**	**9128.37**	**31271.96**

Q1(c)

Weights are adjusted as follows:

$$\mathbf{w}[0] \leftarrow -59.5 + .000002 \times 216.47 = -59.4995$$
$$\mathbf{w}[1] \leftarrow -0.15 + .000002 \times 9128.37 = -0.1317$$
$$\mathbf{w}[2] \leftarrow 0.60 + .000002 \times 31271.96 = 0.6625$$

New predictions are calculated as follows:

```
Prediction(1) = -59.4995 - 0.1317 x 41
                         + 0.6625 x 138 = 26.53
...
...
```

Q1(c)

Calculate the sum of squared errors (i.e. loss) L_2

ID	OxyCon	Age	HeartRate	Prediction	Error	SSE
1	37.99	41	138	26.53	11.46	131.33
2	47.34	42	153	36.33	11.01	121.22
3	44.38	37	151	35.67	8.71	75.86
4	28.17	46	133	22.55	5.62	31.58
5	27.07	48	126	17.65	9.42	88.74
6	37.85	44	145	30.77	7.08	50.13
7	44.72	43	158	39.51	5.21	27.14
8	36.42	46	143	29.18	7.24	52.42
9	31.21	37	138	27.05	4.16	17.31
10	54.85	38	158	40.17	14.68	215.5
11	39.84	43	143	29.57	10.27	105.47
12	30.83	43	138	26.26	4.57	20.88
					Sum	**937.58**
					L_2	468.79

Loss is reducing as you would expect.

Q2

A multivariate logistic regression model has been built to predict the propensity of shoppers to repeat purchase a free gift they are given. The features used by the model are the customer age, the socioeconomic band they belong to (either *a*, *b* or *c*), the average amount of money they spend on each visit and the average number of visits per week they make to the shop. The model is being used by the marketing department to determine who should be given a free gift.

The weights in the trained model are shown in the following table:

Feature	Weight
Intercept (w[0])	-3.83398
Age	-0.02990
Socioeconomic Band B	-0.09089
Socioeconomic Band C	-0.19558
Average Spend	0.02999
Frequency	0.74572

Q2(a)

Explain how there are 6 feature weights in the model given that there are only four original features used.

The intercept is a feature of the model that represents where the fitted line crosses the y-axis.

The socioeconomic band feature is a categorical feature which a regression model cannot handle. It has been converted by one-hot encoding to three features Socioeconomic Band A, Band B and Band C. Band A is not needed as it is implicit in values of zero for Band B and Band C.

Q2(b)

Use this model to make predictions for the following query instances, assume that the positive class is '*yes*':

Query	Age	Socioeconomic Band	Average Spend	Frequency
Q1	56	b	109.32	1.6
Q2	37	c	170.65	0.72
Q3	32	a	165.39	1.08

Model for logistic regression is: $$\mathbb{M}_{\mathbf{w}}(\mathbf{d}) = \frac{1}{1 + e^{-\mathbf{w} \cdot \mathbf{d}}}$$

```
w.q1 = -3.82398
       - .0299 x 56
       - 0.09089 x 1
       + 0.02999 x 109.32
       + 0.74572 x 1.6
       = -1.12
```

$$\mathbb{M}_{\mathbf{w}}(\mathbf{q1}) = \frac{1}{1 + e^{1.12}} = 0.25$$

Value is < 0.5 so prediction is '*no*'

Q2(b)

```
w.q2 = -3.82398
       - 0.0299 x 37
       - 0.19558 x 1
       + 0.02999 x 170.65
       + 0.74572 x 0.72
       = 0.53
```

$$\mathbb{M}_{\mathbf{w}}(\mathbf{q1}) = \frac{1}{1 + e^{-0.53}} = 0.63$$

Value is > 0.5 so prediction is '*yes*'

```
w.q3 = -3.82398
       - 0.0299 x 32
       + 0.02999 x 165.39
       + 0.74572 x 1.08
       = 0.98
```

$$\mathbb{M}_{\mathbf{w}}(\mathbf{q1}) = \frac{1}{1 + e^{-0.98}} = 0.73$$

Value is > 0.5 so prediction is '*yes*'

Q3

A support vector machine has been built to predict whether a patient is at risk of cardiovascular disease. In the dataset used to train the model, the two target levels are *high risk* (the positive class) and *low risk* (the negative class).
The support vectors are shown in the table below where all descriptive variables have been standardised.

SV	Age	BMI	Blood Pressure	Risk
s1	-0.45	0.01	0.22	low risk
s2	-0.28	-0.52	0.36	low risk
s3	0.37	0.09	-1.08	high risk
s4	0.55	0.22	0.21	high risk

In the model the value of w0 is -0.022, and the values of the α parameters are {1.68, 0.23, 0.20, 1.71}

What predictions would this model make for the following query instances?

ID	Age	BMI	Blood Pressure
q1	-0.89	-0.34	0.55
q2	0.45	0.49	-0.47

Q3

Predictions for a support vector machine are:

$$\mathbb{M}(\mathbf{q}) = \sum_{i=1}^{s} (t_i \times \boldsymbol{\alpha}[i] \times (\mathbf{d}_i \cdot \mathbf{q}) + w_0)$$

For q1:

Calculate the dot product of each support vector with q1:

```
q1•s1 = 0.45 x -0.45 + 0.49 x 0.01 - 0.47 x 0.22 = -0.301
q1•s2 = 0.45 x -0.28 + 0.49 x -0.52 - 0.47 x 0.36 = -0.55
q1•s3 = 0.45 x 0.39 + 0.49 x 0.09 - 0.47 x -1.08 = 0.7272
q1•s4 = 0.45 x 0.55 + 0.49 x 0.22 - 0.47 x 0.21 = 0.2566
```

The prediction for the query instance for q1:

```
M(q1) = -1 x 1.68 x (q1•s1) - 0.022
        -1 x 0.23 x (q1•s2) - 0.022
        +1 x 0.02 x (q1•s3) - 0.022
        +1 x 1.71 x (q1•s4) - 0.022
      = 0.9975
```

Value is positive so prediction is '*high risk*'

Q3

For q2:

Calculate the dot product of each support vector with q2:

```
q2•s1 = -0.89 x -0.45 -0.34 x 0.01 + 0.55 x 0.22 = 0.5181
q2•s2 = -0.89 x -0.28 -0.34 x -0.52 + 0.55 x 0.36 = 0.624
q2•s3 = -0.89 x 0.39 -0.34 x 0.09 + 0.55 x -1.08 = -0.9717
q2•s4 = -0.89 x 0.55 -0.34 x 0.22 + 0.55 x 0.21 = -0.4488
```

The prediction for the query instance for q2:

```
M(q2) = -1 x 1.68 x (q1•s1) - 0.022
        -1 x 0.23 x (q1•s2) - 0.022
        +1 x 0.02 x (q1•s3) - 0.022
        +1 x 1.71 x (q1•s4) - 0.022
      = -1.888
```

Value is positive so prediction is '*high risk*'