1. In Lecture 1 we learned about several different types of general data science techniques: (i) classification, (ii) scoring (or regression), (iii) anomaly detection, (iv) clustering, (v) recommending systems and (vi) forecasting. For each of the following problems, suggest which of these methods is most appropriate and justify your selection:

    (a) Discovering hidden patterns in the behaviour of listeners on Spotify? [1 mark]

Clustering. These hidden patterns are not known to the analyst beforehand hence, there is no necessity to guess "a value" from the output. However, the output gives the analyst useful information which can be grouped – for example genre of choice in music of the listeners on Spotify – laying out the options for the analyst to choose which combination benefits most.

    (b) Predicting the life expectancy of a turtle from variables such as its diet and genetic makeup? [1 mark]

Forecasting. The analyst could implement a classifier to differentiate a set of turtles with different diet intake and genetic makeup and then make a prediction of the future by extracting useful information from the processed data. For instance, the turtle with a handicap may find it difficult to fulfil its basic needs leading to shorter life expectancy.

    (c) Determining whether the Apple share price will go up or down over the next month? [1 mark]

Forecasting. Given past events such as lawsuits involvement, Apple share price may be affected; hinting a change in its current trend of stock price maintenance enabling the analyst to predict whether it's going to increase or decrease over the next month.

    (d) Trying to work out whether an image contains a picture of a human running or walking? [1 mark]

Classification. A group of people defines a sample space to be divided into two categories; walking or running based on their guesses.

2. It is common to try and use data collected on consumers or users of websites to try and predict their interests. Imagine we are running a music streaming service, and have information on genres that a user has previously liked or disliked. We could try to use this information to recommend songs from different genres to the user. Table 1 shows the frequency with which users on our service like and dislike two genres of music, Rock and Heavy Metal.

   (a) What is the probability of a person in our population liking Heavy Metal, irrespective of whether they like Rock? [1 mark]

$$let\ HM = Heavy\ Metal$$

$$P(X = HM) = 0.05 + 0.15$$

$$= 0.20$$

   (b) What is the probability of liking Heavy Metal given that a person does not like Rock? [1 mark]

$$let\ HM = Heavy\ Metal\ , nR = not\ Rock$$

$$P(HM|nR) = \frac{0.05}{0.70 + 0.15}$$

$$= \frac{0.05}{0.75}$$

   (c) What is the probability of liking Heavy Metal given a person does like Rock? [1 mark]

$$let\ HM = Heavy\ Metal\ , R = Rock$$

$$P(HM|R) = \frac{0.15}{0.10 + 0.15}$$

$$= \frac{0.15}{0.25}$$

   (d) Do you think liking Rock music is a good predictor of whether a person will like Heavy Metal? Why or why not? [1 mark]

Yes, because the conditional probability to determine whether or not a person likes Heavy Metal given that they like Rock gives a clear difference between the values allowing easy comparison.

3. Imagine that we roll two fair six-sided dice (i.e., all six sides have equal probability). Let X_1 and X_2 be the random variables representing these outcomes. Now, imagine we take one of the dice rolls, say X_1, and add a (possibly negative) constant c to the result. If this becomes less than zero, then we set it to zero; denote this by

$$(X+c)_+ = max(X+c, 0).$$

This type of dice roll manipulation occurs frequently in many board and tabletop games.

(a) What is the expected value of $(X_1+1)_+$? [1 mark]

$$E(X_1+1) = 2\left(\frac{1}{6}\right)+3\left(\frac{1}{6}\right)+4\left(\frac{1}{6}\right)+5\left(\frac{1}{6}\right)+6\left(\frac{1}{6}\right)+7\left(\frac{1}{6}\right)$$

$$= \frac{27}{6}$$

(b) What is the expected value of $(X_1-2)_+$? [1 mark]

$$E(X_1-2) = \frac{1}{6}+2\left(\frac{1}{6}\right)+3\left(\frac{1}{6}\right)+4\left(\frac{1}{6}\right)$$

$$= \frac{10}{6}$$

(c) What is the expected value of $(X_1-2)_+ \times (X_2+1)_+$? [1 mark]

$$E(X_1-2) \times E(X_2+1) = \frac{27}{6}\left(\frac{10}{6}\right)$$

$$= \frac{15}{2}$$

(d) What is the variance of $(X_1-2)_+$? [1 mark]

$$V(X_1-2) = \left[\frac{1}{6}+4\left(\frac{1}{6}\right)+9\left(\frac{1}{6}\right)+16\left(\frac{1}{6}\right)\right]-\left[\frac{1}{6}+2\left(\frac{1}{6}\right)+3\left(\frac{1}{6}\right)+4\left(\frac{1}{6}\right)\right]^2$$

$$= \frac{20}{9}$$

(e) What is the probability that $(X_1-1)_+ > X_2$? [1 mark]

You must show the working/reasoning as to how you obtained these answers

$(X_1-1)_+ / X$	0	1	2	3	4	5
1	-	-	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	-	-	-	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	-	-	-	-	$\frac{1}{36}$	$\frac{1}{36}$
4	-	-	-	-	-	$\frac{1}{36}$
5	-	-	-	-	-	-
6	-	-	-	-	-	-
Total	-	-	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$

$$\therefore \textit{Total probability} = \frac{1}{36}+\frac{2}{36}+\frac{3}{36}+\frac{4}{36}$$

$$= \frac{5}{18}$$

4. Imagine we receive a dataset from a colleague regarding a credit assessment for a bank, and we are asked to model the measurements. What distribution would be appropriate for the following variables (briefly justify your answer):

   (a) Relationship status of the applicant (single or partnered)? [1 mark]

The Bernoulli distribution. The applicant is the random variable in context which can only take on one of the two different values; single or partnered.

   (b) Number of previous times defaulted on loan repayment? [1 mark]

The Poisson distribution models counts of occurrences of things over a time period.

   (c) Income in last financial year? [1 mark]

The Gaussian distribution. The random variable income can take an infinite number of different values over the last 12 months.

   (d) Number of dependents (children, spouse) of applicant? [1 mark]

The Gaussian distribution. The bank could obtain the probability of applicants having n number of dependents.

5. Imagine that a continuous random variable X defined on the range $[0, b]$ follows the probability density function

$$p(X=x|b)=\begin{cases}\frac{2x}{b^2} \text{ for } x \epsilon [0,b] \\ 0 \text{ everywhere else}\end{cases}$$

Answer the following questions; you must include working if appropriate.

(a) Determine the expected value of X, i.e., $E[X]$. [1 mark]

$$E(x)=\int_0^b x\left(\frac{2x}{b^2}\right)dx$$

$$=\frac{2}{b^2}\int_0^b x^2\,dx$$

$$=\frac{2}{b^2}\left[\frac{x^3}{3}\right]b0$$

$$=\frac{2b}{3}$$

(b) Determine the cumulative distribution function for this distribution, i.e., $P(X \le x)$. [1 mark]

$$P(X\le x)=\int_0^x \frac{2x}{b^2}dx$$

$$=\frac{2}{b^2}\int_0^x x\,dx$$

$$=\frac{2}{b^2}\left[\frac{x^2}{2}\right]x0$$

$$=\frac{x^2}{b^2}$$

(c) What is the median value of this distribution? [1 mark]

$$\frac{median^2}{b^2}=\frac{1}{2}$$

$$median=\sqrt{\frac{b^2}{2}}$$

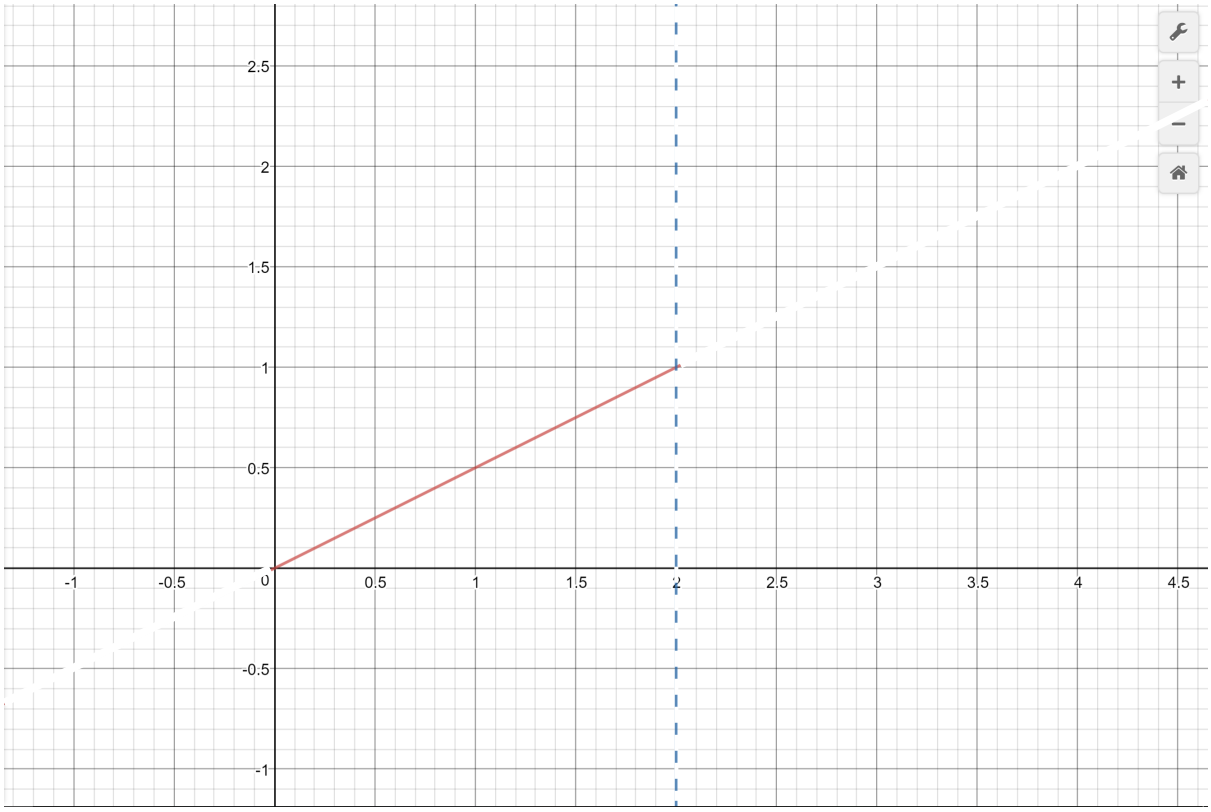
(d) Plot the probability density function of X when b = 2. [1 mark]

$$y=\frac{2x}{2^2} \text{ for } x \epsilon [0,2]$$

$$When\ x=0,\ y=0$$

$$When\ x=1,\ y=\frac{1}{2}$$

$$When\ x=2,\ y=1$$



6. A clothes store records the heights of a number of customers buying trousers in the previous week. The recorded heights (in metres) were

$$y=(1.78,1.65,1.62,1.84,1.75,1.85,1.52,1.55).$$

The stock manager decides to use this information to help determine how much future product to order. To do so, she decides to fit a normal distribution to this data and use it to model the population of future people buying trousers. You must show working/R code as required to obtain full marks.

(a) Fit a normal distribution to the data y using the maximum likelihood estimator for μ and σ. What are the values of these parameters for this data? [2 marks]

```
> mean = (1.78 + 1.65 + 1.62 + 1.84 + 1.75 + 1.85 + 1.52 + 1.55)/8
> mean
[1] 1.695

> sd = sqrt(((1.78 - mean)^2 + (1.65 - mean)^2 + (1.62 - mean)^2 + (1.84 - mean)^2 + (1
.75 - mean)^2 + (1.85 - mean)^2 + (1.52 - mean)^2 + (1.55 - mean)^2)/8)
> sd
[1] 0.1196871
```

(b) Plug these estimates $\hat{\mu}$ and $\hat{\sigma}$ into the normal distribution, and use this to make predictions about future customers. Using this model, answer the following questions:

i. Imagine the store stocks pants suitable for people in the following four height ranges:

$$(¿1.5m),(1.5m-1.65m),(1.65m-1.8m),(¿1.8m)$$

What are the estimated proportions of people in the population of future customers that would fall into each of these height ranges? [2 marks]

```
> pnorm(1.5, mean, sd)
[1] 0.05163023

> pnorm(1.65, mean, sd) - pnorm(1.5, mean, sd)
[1] 0.3018355

> pnorm(1.8, mean, sd) - pnorm(1.65, mean, sd)
[1] 0.456369

> 1 - pnorm(1.8, mean, sd)
[1] 0.1901652
```

ii. If a new customer walks into the store to buy pants, which height range are they most likely to be in? [1 mark]

1.65m – 1.8m because the probability of a new customer to be in between 1.65m and 1.8m is highest.

iii. If the store receives 10 customers in one day, what is the probability that none of them will be 1.65m or taller? [1 mark]

```
> dbinom(10, 10, pnorm(1.65, mean, sd))
[1] 3.0442e-05
```

iv. Imagine the store receives 160 customers per week buying trousers. During a week of sales how many pairs of pants for people between 1.65m and 1.8m would the store expect to sell? [1 mark]

```
> 160*(pnorm(1.8, mean, sd)-pnorm(1.65, mean, sd))
[1] 73.01905
```