FIT2086 Assignment 1
Lim Zheng Haur 32023952

Question 1

1. Predicting the amount of bananas that will purchased in the next week? [1 mark]
   Forecasting

2. Determining whether someone's credit card has been compromised via purchasing history? [1 mark]
   Anomaly detection

3. Discovering the genre preferences of Netflix users? [1 mark]
   Recommendation systems

4. Using genomic and lifestyle factors to determine the chance of an individual contracting breast cancer in the next year? [1 mark]
   Risk prediction

Question 2

1. Using the frequencies provided above fill in the entries of Table 1 with the proportions of the times those events occurred, i.e., estimates of the joint probabilities of a win/draw/loss for home/away games (up to 3 decimal places). [1 mark]

	R = 0	**R = 1**	**R = 2**
H = 0	**0.053**	**0.211**	**0.237**
H = 1	**0.132**	**0.053**	**0.316**

2. Using these proportions, calculate the marginal probability of Barcelona winning a game, regardless of whether it is played at home or away, i.e., P(R = 2). [1 mark]
   P(R = 2) = 0.237 + 0.316
   P(R = 2) = 0.553

3. What is the probability that Barcelona will win a game, given that they are playing at home? [1 mark]
   P(R = 2 | H = 1) = 0.316 / 0.500
   P(R = 2 | H = 1) = 0.632

4. What is the probability that Barcelona will win a game, given that they are playing away? [1 mark]
   P(R = 2 | H = 0) = 0.237 / 0.500
   P(R = 2 | H = 0) = 0.474

5. Do you believe that Barcelona is more likely to win when at home versus when they play away? [1 mark]
   I believe they are more likely to win when at home than at away.
   P(R = 2 | H = 1) = 0.632 while P(R = 2 | H = 0) = 0.474. The probability of winning when at home is higher than at away.

6. Imagine that Barcelona will play an away game, then a home game, and then an away game in their next three games. What is the probability that they will not lose two out of three of these games? [2 marks]
   The different possibilities losing exactly 2 of the 3 games in the sequence "away home away" are:
   Lose Lose Win/Draw = P(R=0 | H=0) x P(R=0 | H=1) x (P(R=2 | H=0) + P(R=1 | H=0))
   = 0.105 x 0.263 x (0.474 + 0.421)
   = 0.025
   Lose Win/Draw Lose = P(R=0 | H=0) x (P(R=2 | H=1) + P(R=1 | H=1)) x P(R=0 | H=0)
   = 0.105 x (0.632 + 0.105) x 0.105
   = 0.008
   Win/Draw Lose Lose = (P(R=2 | H=0) + P(R=1 | H=0)) x P(R=0 | H=1) x P(R=0 | H=0)
   = (0.474 + 0.421) x 0.263 x 0.105
   = 0.025

   Probability of losing 2 out of 3 games = 0.025 + 0.008 + 0.025 = 0.058
   Probability of not losing 2 out of 3 games = 1 – 0.058

   Therefore, Probability of not losing 2 out of 3 games = 0.942

Question 3

1. What is the variance of S, i.e., what is V [S]? [1 mark]

$$\boldsymbol{V[S] = V[X] + \mathrm{V}[3\mathrm{Y}]}$$
$$\boldsymbol{V[S] = 2.917 + 11.250}$$
$$\boldsymbol{V[S] = 14.167}$$

2. Determine the probability distribution of S, i.e., the probability that S ∈ {4, . . . , 18}. [1 mark]

S	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
p(S)	**1/24**	**1/24**	**1/24**	**2/24**	**2/24**	**2/24**	**2/24**	**2/24**	**2/24**	**2/24**	**2/24**	**2/24**	**1/24**	**1/24**	**1/24**

3. What is the expected value of √ S, i.e., what is E [√ S]? [1 mark]

S	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
p(S)	**1/24**	**1/24**	**1/24**	**2/24**	**2/24**	**2/24**	**2/24**	**2/24**	**2/24**	**2/24**	**2/24**	**2/24**	**1/24**	**1/24**	**1/24**
√S	**2.000**	**2.236**	**2.449**	**2.646**	**2.828**	**3.000**	**3.162**	**3.317**	**3.464**	**3.606**	**3.742**	**3.873**	**4.000**	**4.123**	**4.243**
√S p(S)	**0.083**	**0.093**	**0.102**	**0.220**	**0.236**	**0.250**	**0.264**	**0.276**	**0.289**	**0.300**	**0.312**	**0.323**	**0.167**	**0.172**	**0.177**

$$\boldsymbol{E\left[\sqrt{S}\right] = \sum \sqrt{S}\, p(S) = 3.264}$$

4. Calculate the approximate value of E [√ S] using the Taylor-series procedure discussed in Lecture 2. [2 marks]

$$\boldsymbol{\frac{d^2\sqrt{E[S]}}{dx^2} = -\frac{1}{4\sqrt{E[S]^3}}}$$
$$\boldsymbol{E\left[\sqrt{S}\right] \approx \sqrt{E[S]} + \left[-\frac{1}{4\sqrt{E[S]^3}}\right]\frac{V[S]}{2}}$$
$$\boldsymbol{E\left[\sqrt{S}\right] \approx \sqrt{11} + \left[-\frac{1}{4\sqrt{11^3}}\right]\frac{14.167}{2}}$$
$$\boldsymbol{E\left[\sqrt{S}\right] \approx 3.268}$$

5. Imagine that we roll a second fair four-sided die; call the outcome of this roll Y2. What is the expected value of $(X_1 + 3Y_1 - 2Y_2)^2$, i.e., what is $E[(X_1 + 3Y_1 - 2Y_2)^2]$? [2 marks]

$$\boldsymbol{let\ S = X_1 + 3Y_1 - 2Y_2}$$

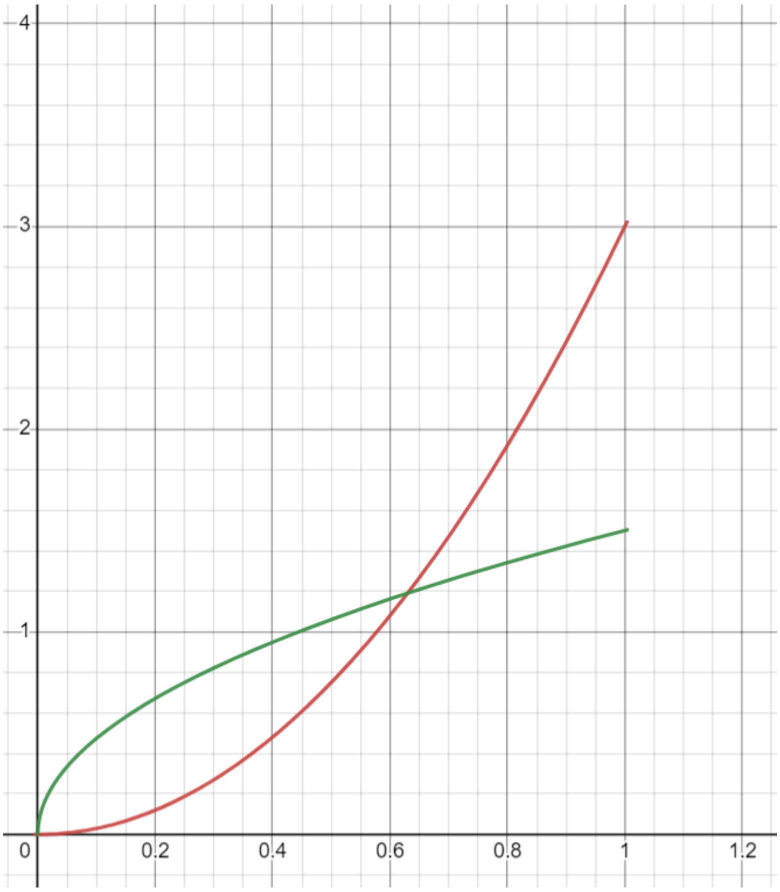
S	-4	-3	-2	-1	0	1	2	3	4	5	6
p(S)	1/96	1/96	2/96	3/96	4/96	5/96	6/96	7/96	7/96	8/96	8/96
S^2	16	9	4	1	0	1	4	9	16	25	36
S^2p(S)	0.167	0.094	0.083	0.031	0.000	0.052	0.250	0.656	1.167	2.083	3.000

S	7	8	9	10	11	12	13	14	15	16	
p(S)	8/96	7/96	7/96	6/96	5/96	4/96	3/96	2/96	1/96	1/96	
S^2	49	64	81	100	121	144	169	196	225	256	
S^2p(S)	4.083	4.667	5.906	6.250	6.302	6.000	5.281	4.083	2.344	2.667	

$$\boldsymbol{E\left[(X_1 + 3Y_1 - 2Y_2)^2\right] = 55.167}$$

Question 4

1. Plot the probability density function of X when a = 1/2 and a = 2 for x ∈ [0, 1]. [2 marks]
   Green when a = 1/2
   Red when a = 2



2. Determine the expected value of X, i.e., E [X]. [1 mark]

$$\boldsymbol{E[X] = \int xp(x)dx}$$

$$\boldsymbol{E[X] = \int_0^1 x[(a+1)x^a]dx}$$

$$\boldsymbol{E[X] = (a+1)\int_0^1 x^{a+1}dx}$$

$$\boldsymbol{E[X] = (a+1)\left[\frac{x^{a+2}}{a+2}\right]_0^1}$$

$$\boldsymbol{E[X] = \frac{a+1}{a+2}}$$

3. Determine the expected value of 1/X, i.e., E [1/X]. [1 mark]

$$E\left[\frac{1}{X}\right] = \int \frac{1}{x} p(x)dx$$

$$E\left[\frac{1}{X}\right] = \int_0^1 x^{-1}[(a+1)x^a]dx$$

$$E\left[\frac{1}{X}\right] = (a+1)\int_0^1 x^{a-1}dx$$

$$E\left[\frac{1}{X}\right] = (a+1)\left[\frac{x^a}{a}\right]_0^1$$

$$E\left[\frac{1}{X}\right] = \frac{a+1}{a}$$

4. Determine the variance of X, i.e., V [X]. [1 mark]

$$V[X] = E[X^2] - E[X]^2$$

$$V[X] = \int_0^1 x^2[(a+1)x^a]dx - E[X]^2$$

$$V[X] = (a+1)\left[\frac{x^{a+3}}{a+3}\right]_0^1 - \left(\frac{a+1}{a+2}\right)^2$$

$$V[X] = \frac{a+1}{a+3} - \frac{a^2+2a+1}{a^2+4a+4}$$

$$V[X] = \frac{(a+1)(a+2)^2}{(a+3)(a+2)^2} - \frac{(a+1)^2(a+3)}{(a+3)(a+2)^2}$$

$$V[X] = \frac{a+1}{(a+3)(a+2)^2}$$

5. Determine the median of X. [2 marks]

$$\int_0^{med} [(a+1)x^a]dx = \frac{1}{2}$$

$$(a+1)\left[\frac{x^{a+1}}{a+1}\right]_0^{med} = \frac{1}{2}$$

$$med^{a+1} = \frac{1}{2}$$

$$med = 2^{\frac{-1}{a+1}}$$

Question 5

```
> dog_bites <- read.csv("dogbites.1997.csv")
```

1. Fit a Poisson distribution to the dog bites data using maximum likelihood. What is the estimated rate, λˆ, of dog bite incidents in Australia during this period? [1 marks]

   ```
   > lambda <- mean(dog_bites$daily.dogbites)
   > lambda
   [1] 4.391534
   > # estimated rate of dog bites per day = 4.391534
   ```

2. Plug the estimated λˆ into the Poisson distribution, and use this to make predictions about future dog bite incidences. Using this model, answer the following questions:

   a. What is the probability of two or less admissions for a dog-bite in a day? [1 mark]

      ```
      > less_than_2 <- ppois(2, lambda)
      > less_than_2
      [1] 0.1861507
      > # probability of two or less dog bites = 0.1861507
      ```

   b. What are the two most likely number of dog-bite admissions to occur on any given day? [1 mark]

      ```
      > dpois_dog_bites <- dpois(0:max(dog_bites), lambda)

      > # finding 2 highest probability
      > highest <- sort(dpois_dog_bites)[length(dpois_dog_bites)]
      > highest2 <- sort(dpois_dog_bites)[length(dpois_dog_bites)-1]

      > # finding the index of the 2 highest probability
      > which(dpois_dog_bites == highest)-1
      [1] 4
      > which(dpois_dog_bites == highest2)-1
      [1] 3
      > # the two most likely number of dog bites = 4 and 3
      ```

   c. Over a week period (i.e., 7 days), what is the probability of seeing at most 32 dogbites? [1 mark]

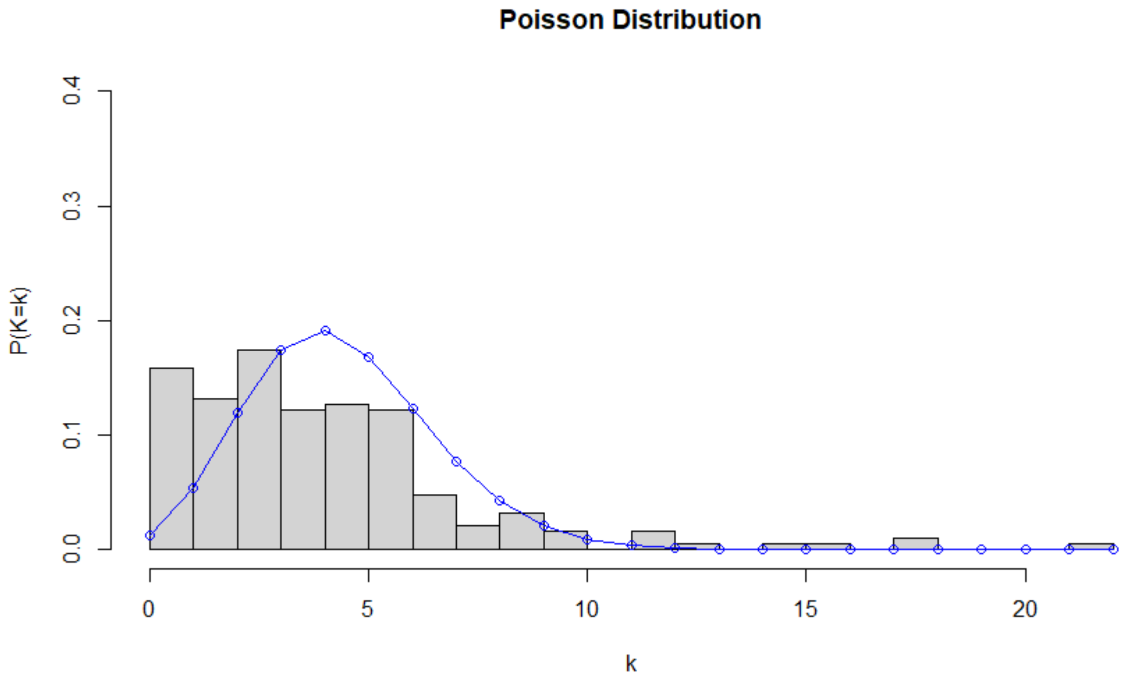
      ```
      > less_than_32 <- ppois(32, lambda*7)
      > less_than_32
      [1] 0.6346646
      > # probability of at most 32 dog bites in a week = 0.6346646
      ```

d. What is the probability of seeing three or more dog-bite admissions for at least 12 days in a 14 day period? [1 mark]

```
> more_than_3 <- 1 - ppois(2, lambda)
> less_than_12_of_14 <- pbinom(11, 14, more_than_3)
> more_than_12_of_14 <- 1 - less_than_12_of_14
> more_than_12_of_14
[1] 0.5012691
> # probability of more than 3 dog bite at least 12 of 14 days
= 0.5012691
```

3. The quality of predictions of a model are only as good as the model is itself representative of the population. Do you believe that the Poisson distribution is an appropriate model for the dog bite data? Plot the observed probabilities of the different number of daily dog-bite incidences against the probabilities for the number of daily dog-bites as predicted by your Poisson model (over the range of 0 to 22 dog-bite admissions), and use this to justify whether or not this model is a good fit to the data. [2 marks]

```
> # plotting a histogram
> hist(dog_bites$daily.dogbites, breaks = 22, freq = F, xlim =
c(0,22), ylim = c(0,0.4), xlab = "k", ylab = "P(K=k)", main =
"Poisson Distribution")
> # adding the poisson distribution line
> xpois = seq(from = 0, to = 20, by = 1)
> lines(x = xpois, y = dpois(x = xpois, lambda), type = "l", col =
"blue")
> lines(x = xpois, y = dpois(x = xpois, lambda), type = "p", col =
"blue")
```



Based on the diagram above, the actual observed number or daily incidences generally follows the predictions of the Poisson Distribution. Hence, it shows that Poisson model is a good fit of the data.