CSCI 347
Homework 02

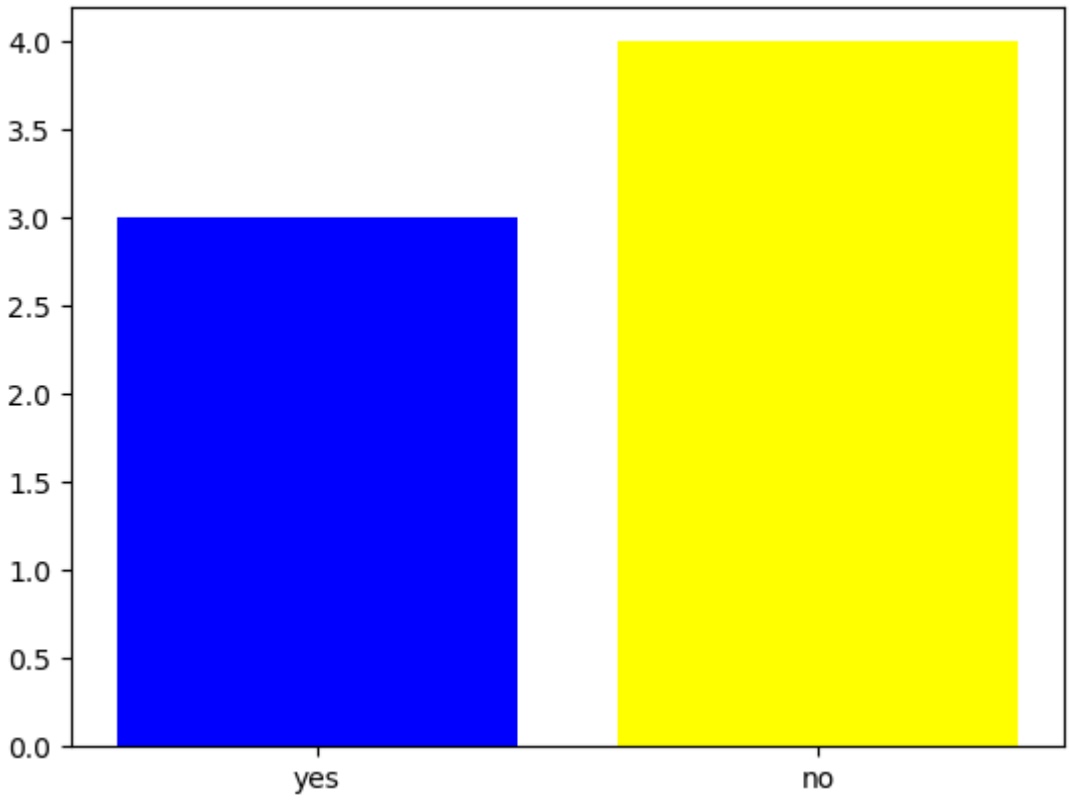
Show your work. Include any code snippets you used to generate an answer, using comments in the code to clearly indicate which problem corresponds to which code

Consider the following data matrix

$$D = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{ccc} X_1 & X_2 & X_3 \\ red & yes & north \\ blue & no & south \\ yellow & no & east \\ yellow & no & west \\ red & yes & north \\ yellow & yes & north \\ blue & no & west \end{array}$$

Answer the following:

1. (5 points) Use matplotlib to create a bar plot for the counts of the variable X_2. Make sure to label the axis.



```
import numpy as np
import matplotlib.pyplot as plt

d = np.array([['red', 'yes', 'north'], ['blue', 'no', 'south'],
              ['yellow', 'no', 'east'], ['yellow', 'no', 'west'],
              ['red', 'yes', 'north'], ['yellow', 'yes', 'north'],
              ['blue', 'no', 'west']])

X2 = d[:,1]
countY = sum(X2 == 'yes')
countN = sum(X2 == 'no')

plt.bar(x = ('yes','no'), height = (countY, countN),
        color = ('blue', 'yellow'))
plt.show()
```

2. (2 points) Use one-hot encoding to transform all the categorical attributes to numerical values. Write down the transformed data matrix. (In what follows, we will referred to the transformed data matrix as Y).

$$Y = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{ccccccccc} X_1r & X_1b & X_1y & X_2y & X_2n & X_3n & X_3e & X_3s & X_3w \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}$$

3. (2 points) What is the Euclidean distance between instance x_2 (second row) and x_7 (seventh row) after applying one-hot encoding.
$\sqrt{(0-0)^2+(1-1)^2+(0-0)^2+(0-0)^2+(1-1)^2+(0-0)^2+(0-0)^2+(1-0)^2+(0-1)^2}$
$= \sqrt{2} = \mathbf{1.4142}$

4. (2 points) What is the cosine similarity (cosine of the angle) between data instance x_2 and data instance x_7 after applying one-hot encoding?
$x_2 \cdot x_7 = 0*0+1*1+0*0+0*0+1*1+0*0+0*0+1*0+0*1 = 2$
$||x_2|| = \sqrt{(0)^2+(1)^2+(0)^2+(0)^2+(1)^2+(0)^2+(0)^2+(1)^2+(0)^2} = \sqrt{3}$
$||x_7|| = \sqrt{(0)^2+(1)^2+(0)^2+(0)^2+(1)^2+(0)^2+(0)^2+(0)^2+(1)^2} = \sqrt{3}$
$\frac{2}{\sqrt{3}*\sqrt{3}} = \mathbf{\frac{2}{3}}$

5. (2 points) What is the Hamming distance between data instance x_2 and data instance x_7 after applying one-hot encoding?
   $0 \bigoplus 0+1 \bigoplus 1+0 \bigoplus 0+0 \bigoplus 0+1 \bigoplus 1+0 \bigoplus 0+0 \bigoplus 0+1 \bigoplus 0+0 \bigoplus 1$
   $= 0+0+0+0+0+0+0+1+1 = \mathbf{2}$

6. (2 points) What is the Jaccard similarity between data instance x_2 and x_7 after applying one-hot encoding?
   $\frac{0 \bigwedge 0+1 \bigwedge 1+0 \bigwedge 0+0 \bigwedge 0+1 \bigwedge 1+0 \bigwedge 0+0 \bigwedge 0+1 \bigwedge 0+0 \bigwedge 1}{0 \bigvee 0+1 \bigvee 1+0 \bigvee 0+0 \bigvee 0+1 \bigvee 1+0 \bigvee 0+0 \bigvee 0+1 \bigvee 0+0 \bigvee 1}$
   $= \frac{0+1+0+0+1+0+0+0+0}{0+1+0+0+1+0+0+1+1} = \mathbf{\frac{1}{2}}$

7. (2 points) What is the multivariate mean of Y?

```
import pandas as pd

data =  [[1, 0, 0, 1, 0, 1, 0 ,0 ,0],
         [0, 1, 0, 0, 1, 0, 0, 1, 0],
         [0, 0, 1, 0, 1, 0, 1, 0, 0],
         [0, 0, 1, 0, 1, 0, 0, 0, 1],
         [1, 0, 0, 1, 0, 1, 0, 0, 0],
         [0, 0, 1, 1, 0, 1, 0, 0, 0],
         [0, 1, 0, 0, 1, 0, 0, 0, 1]]

df = pd.DataFrame(data)

x = df.mean()
print(x)
```

   multivariate mean = $\mathbf{0.285714}, \mathbf{0.285714}, \mathbf{0.428571}, \mathbf{0.428571}, \mathbf{0.571429}$, $\mathbf{0.428571}, \mathbf{0.142857}, \mathbf{0.142857}, \mathbf{0.285714}$

8. (2 points) What is the estimated variance of the first column of Y?
   $\frac{1}{7}(1+0+0+0+1+0+0) = 0.286$
   $\frac{1}{6}(1-0.286)^2+(0-0.286)^2+(0-0.286)^2+(0-0.286)^2+(1-0.286)^2+(0-0.286)^2+(0-0.286)^2)$
   $= \mathbf{0.238}$

9. (2 points) What is the resulting matrix after applying standard (z-score) normalization to the matrix Y. In the following, we will call this matrix Z.

```
import numpy as np

data =  [[1, 0, 0, 1, 0, 1, 0 ,0 ,0],
         [0, 1, 0, 0, 1, 0, 0, 1, 0],
         [0, 0, 1, 0, 1, 0, 1, 0, 0],
         [0, 0, 1, 0, 1, 0, 0, 0, 1],
         [1, 0, 0, 1, 0, 1, 0, 0, 0],
         [0, 0, 1, 1, 0, 1, 0, 0, 0],
         [0, 1, 0, 0, 1, 0, 0, 0, 1]]

myArray = np.array(data)

mean = np.mean(myArray, axis = 0)
std = np.std(myArray, axis = 0)
print(mean)
print("-----")
print(std)
print("----------------")
Z = ((myArray-mean)/std)
print(Z)
```

$$Z = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{c} X_1r \\ 1.58 \\ -0.63 \\ -0.63 \\ -0.63 \\ 1.58 \\ -0.63 \\ -0.63 \end{array} \begin{array}{c} X_1b \\ -0.63 \\ 1.58 \\ -0.63 \\ -0.63 \\ -0.63 \\ -0.63 \\ 1.15 \end{array} \begin{array}{c} X_1y \\ -0.86 \\ -0.86 \\ 1.15 \\ 1.15 \\ -0.86 \\ 1.15 \\ -0.86 \end{array} \begin{array}{c} X_2y \\ 1.15 \\ -0.86 \\ -0.86 \\ -0.86 \\ 1.15 \\ 1.15 \\ -0.86 \end{array} \begin{array}{c} X_2n \\ -1.15 \\ 0.86 \\ 0.86 \\ 0.86 \\ -1.15 \\ -1.15 \\ 0.86 \end{array} \begin{array}{c} X_3n \\ 1.15 \\ -0.86 \\ -0.86 \\ -0.86 \\ 1.15 \\ 1.15 \\ -0.86 \end{array} \begin{array}{c} X_3e \\ -0.41 \\ -0.41 \\ 2.45 \\ -0.41 \\ -0.41 \\ -0.41 \\ -0.41 \end{array} \begin{array}{c} X_3s \\ -0.41 \\ 2.45 \\ -0.41 \\ -0.41 \\ -0.41 \\ -0.41 \\ -0.41 \end{array} \begin{array}{c} X_3w \\ -0.63 \\ -0.63 \\ -0.63 \\ 1.58 \\ -0.63 \\ -0.63 \\ 1.58 \end{array}$$

10. (2 points) What is the multivariate mean of Z?

```
Zmean = np.mean(Z, axis = 0)
print(Zmean)
```

[3.17206578e-17 3.17206578e-17 3.17206578e-17 1.58603289e-174.75809868e-17 1.58603289e-1
6.34413157e-17 3.17206578e-17 3.17206578e-17] = [0 0 0 0 0 0 0 0 0]

11. (2 points) Let z_i be the i-th row of Z. What is Euclidean distance between z_2 and z_7?

```
temp = Z[1] - Z[6]
sumsq = np.dot(temp.T, temp)
print(np.sqrt(sumsq))
```

= **3.614784**

Acknowledgements: Homework problems adapted from assignments of Veronika Strnadova-Neeley.