Questions

1. There are 150 students in our B565 class. Assume each student has a 2% of chance of carrying coronavirus. We also know that the Omicron variants dominate and account for most of the new cases (95%). What's the probability that the entire class is free of coronavirus and Omicron variants, respectively? Show your work.

Answer -

P(Corona) = 0.02 * 150 = 3 students
P(Omicron) = 0.95 * 3 = 2.85 students
P(free from corona) = ((150-3)/150) * 100 = 98%
P(free from Omicron) = ((150-2.85)/150) * 100 = 98.1%

2. You are given a set of m objects that is divided into G groups, where the i group is of size mi. If the goal is to obtain a sample of size n <m, what is the difference between the following two sampling schemes? (Assume sampling with replacement is used). Show a couple of scenarios (or analysis goals), in which you will prefer one strategy over the other. [15 pts]
(a) You randomly select n ×mi /m elements from each group.
(b) You randomly select n elements from the data set, without regard for the group to which an object belongs.

Answer -

In the sampling strategy 'a' the sample from each group is proportional to its size relative to the total number of objects. The second scheme (b) is a simple random sampling. When we use the 'a' sampling scheme we are assured of a sample from each group, which can be used to estimate various statistical parameters of that group whenever the corresponding sample size is large enough and the sample is homogeneous.The variance in scheme 'a' is lower than scheme 'b' as scheme 'b' has to consider variance between the different groups. Therefore I will prefer scheme 'a' if the data is homogeneous.

3. Recall that given two vectors x and y of n dimensions, their cosine similarity, Euclidean distance and correlation can be computed as following. [15 pts]
Sets can also be represented as vectors of zeros and ones, so for those vectors, Jaccard similarity (intersection over union) can be used.For the following vectors x and y, calculate the indicated similarity or distance measures. Show the steps.
• x = (1, 0, 0, 1, 1, 1), y = (1, 1, 1, 0, 0, 1) cosine, correlation, Euclidean, Jaccard
• x = (1, −2, 0, 2, 0, −3), y = (−1, 2, −1, 0, 0, −1) cosine, correlation, Euclidean

Answer -

A) x = (1, 0, 0, 1, 1, 1), y = (1, 1, 1, 0, 0, 1)
   Cosine -
   x . y = 1*1 + 0 * 1 + 0 * 1 + 1 * 0 + 1 * 0 + 1 * 1 = 2
   ||x|| = (1^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2) ^ 0.5 = 2
   ||y|| = (1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2) ^ 0.5 = 2

Cosine Similarity - 2 / ( 2 * 2) = 0.5

Correlation -
$\overline{x}$ = (1+0+0+1+1+1)/6 = 4/6 = 0.67
$\overline{y}$ = (1+1+1+0+0+1)/6 = 4/6 = 0.67

| x | y | X - $\overline{x}$ | Y - $\overline{y}$ |
|---|---|---|---|
| 1 | 1 | 0.33 | 0.33 |
| 0 | 1 | -0.67 | 0.33 |
| 0 | 1 | -0.67 | 0.33 |
| 1 | 0 | 0.33 | -0.67 |
| 1 | 0 | 0.33 | -0.67 |
| 1 | 1 | 0.33 | 0.33 |

Corr = ( 0.33 * 0.33 + (-0.67)*(0.33) + (-0.67)*(0.33) + (0.33) * (-0.67) + (0.33) * (-0.67) + (0.33) * (-0.67) )

---

$\sqrt{(0.03)^2 + (-0.67)^2 + (-0.67)^2 + (0.03)^2 + (0.03)^2 + (0.03)^2}$ x
$\sqrt{(0.03)^2 + (0.03)^2 + (0.03)^2 + (-0.67)^2 + (-0.67)^2 + (0.03)^2}$

Correlation = -0.667 / $\sqrt{(1.3333 * 1.3333)}$ = - 0.5

Euclidean distance -

Euclidean Dist = $\sqrt{(1-1)^2 + (0-1)^2 + (0-1)^2 + (1-0)^2 + (1-0)^2 + (1-1)^2}$

Euclidean Dist = 2

Jaccard -

| $x \cap y$ | = 6
| $x \cup y$ | = 2

Jaccard similarity = | $x \cap y$ | / | $x \cup y$ | = 6 / 2 = 3

B) x = (1, −2, 0, 2, 0, −3), y = (−1, 2, −1, 0, 0, −1)

Cosine -

$x \cdot y$ = 1*-1 + -2*2 + 0*-1 + 2*0 + 0*0 + -3*-1 = -2

||x|| = (1^2 + -2^2 + 0^2 + 2^2 + 0^2 + -3^2) = 18

||y|| = (-1^2 + 2^2 + -1^2 + 0^2 + 0^2 + -1^2) = 7

Cosine Similarity = (-2) / (18 * 7) = -0.0159

Correlation -

$\overline{x}$ = (1-2+0+2+0-3)/6 = -2/6 = -3

$\overline{y}$ = (-1+2-1+0+0-1)/6 = -1/6 = -0.167

| x | y | X - $\overline{x}$ | Y - $\overline{y}$ |
|---|---|---|---|
| 1 | -1 | 4 | -0.833 |
| -2 | 2 | -5 | 2.167 |
| 0 | -1 | -3 | -0.833 |
| 2 | 0 | -1 | 0.167 |
| 0 | 0 | -3 | 0.167 |
| -3 | -1 | 0 | -0.833 |

Corr = -3.322 + (-10.835) + 2.499 + (-0.167) + (-0.501) + 0 / (

$\sqrt{4^2 + (-5)^2 + (-3)^2 + (-1)^2 + (-3)^2 + 0}$x(

$\sqrt{(-0.833)^2 + (2.167)^2 + (-0.833)^2 + (0.167)^2 + (0.167)^2 + (-0.833)^2}$)

Corr = (-12.326) / (7.75) x (2.61) = - 0.60

Euclidean distance -

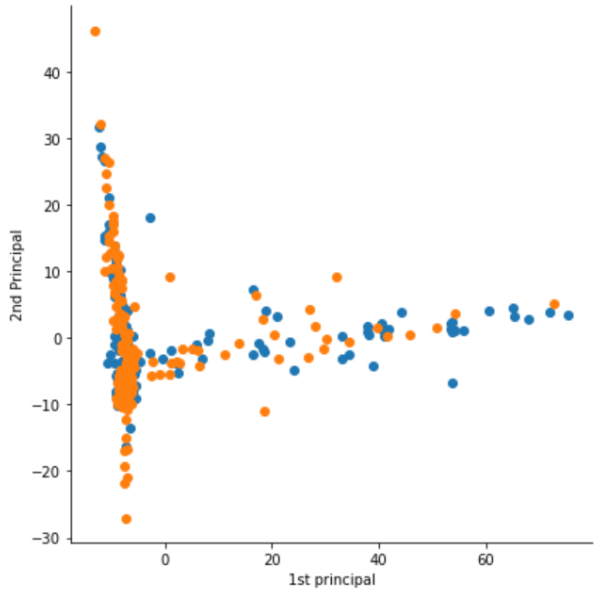
Euclidean Dist = $\sqrt{(1+1)^2 + (-2-2)^2 + (0+1)^2 + (2-0)^2 + (0-0)^2 + (-3+1)^2}$

Euclidean Dist = 5.38

4. We can see that PCA is not really good at differentiating between different patients. Not much data is being captured in PCA and therefore its not a good approach for this dataset. We also can only see 1 huge cluster and then some sparse clusters.
We can see that t-SNE on the other hand performs better than PCA at capturing different clusters of different patients. t-SNE results in a good dimensionality reduction of this dataset since it is a non linear method and adapts to the data.

```
:  1 PCA_data = np.vstack((PCA_data.T,y)).T
   2 pca_df = pd.DataFrame(data=PCA_data,columns=("1st principal", "2nd Principal", 'label'))
   3 sns.FacetGrid(pca_df, hue='label', size = 6).map(plt.scatter, '1st principal', '2nd Principal')
   4 plt.show()
```

```
C:\Users\yashh\anaconda3\lib\site-packages\seaborn\axisgrid.py:337: UserWarning: The `size` parameter has been re
ht`; please update your code.
  warnings.warn(msg, UserWarning)
```




PCA

In [15]:

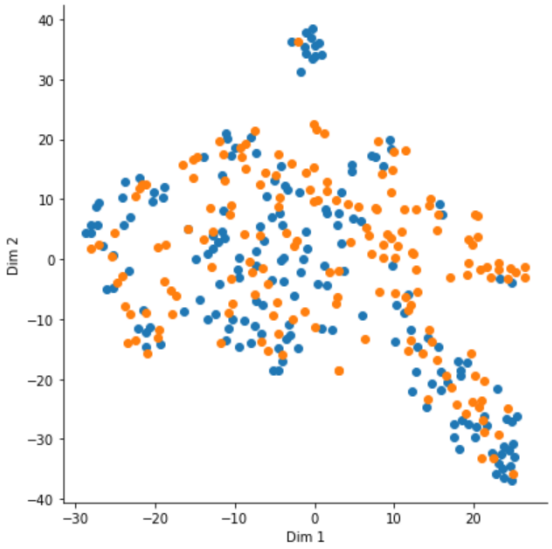
```
X_embedded = np.vstack((X_embedded.T,y)).T

tsne_df = pd.DataFrame(data = X_embedded, columns=("Dim 1", "Dim 2", "label"))

sns.FacetGrid(tsne_df, hue = "label",size =6).map(plt.scatter,'Dim 1','Dim 2')

plt.show()
```

```
C:\Users\yashh\anaconda3\lib\site-packages\seaborn\axisgrid.py:337: UserWarning: The `size` parameter has been renar
ht`; please update your code.
  warnings.warn(msg, UserWarning)
```




t-SNE