1. We know that there is a 2% chance of anyone getting Corona, so the probability of no one getting Corona is 98%. For 160 people, the probability would be $(0.98)^{160} = 0.039$

Now to calculate the probability of not getting Eris, we find the probability of the student being covid and eris free. Which is 0.98 * (1-0.17) = 0.8134. The probability for 160 students will be $(0.8134)^{160} = 4.45 * 10^{-15}$.

2. There are 2 approaches being considered here: first one is weighted selection and the other is totally random selection. Now when you consider both the cases, the difference between the two approaches is that the first one considers the total group size and the second one does not. The difference in the degree of independence offered in both cases results in a bias in the selection for the second case. Whereas the first case considers equal representation of all categories.

   This can be represented with a case. Consider America's political system that is divided into two parties: republicans and democrats. The voter pool would consist of republican and democrat voters. Presumably in a vastly unequal proportion per state. If we consider the first approach for this situation, it would be more beneficial if our goal was to study the respective voter parties equally without considering how many total voters of that party exist. However, if we want a more independent approach that seeks to evaluate the impact of each party's vote on to the party's success. We can utilize the second approach that disregards the total category population.

   Thus both approaches work for this situation.

3.

Cosine similarity, $\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}}{\|\mathbf{x}\|} \cdot \frac{\mathbf{y}}{\|\mathbf{y}\|}$

Euclidean distance, $d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y})}$

Correlation, $corr(\mathbf{x}, \mathbf{y}) = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

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 $\mathbf{x}$ and $\mathbf{y}$, calculate the indicated similarity or distance measures. Show the steps.

- $\mathbf{x} = (1,0,1,1,1)$, $\mathbf{y} = (1,1,0,1,1)$ cosine, correlation, Euclidean, Jaccard

- $\mathbf{x} = (1, -2, 2, -3),\ \mathbf{y} = (-1, 2, -3, -1)$ cosine, correlation, Euclidean

(a) x = (1,0,1,1,1), y = (1,1,0,1,1)

cos(x,y) = $\frac{1+0+0+1+1}{\sqrt{1^2+0^2+1^2+1^2+1^2} * \sqrt{1^2+1^2+0^2+1^2+1^2}} = \frac{3}{\sqrt{4} * \sqrt{4}} = \frac{3}{4} = 0.75$

Euclidean distance, d(x,y) =

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} = \sqrt{(1-1)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 + (1-1)^2} = \sqrt{2} = 1.414$

corr(x,y) =

$\frac{(1-0.8)(1-0.8) + (0-0.8)(1-0.8) + (1-0.8)(0-0.8) + (1-0.8)(1-0.8) + (1-0.8)(1-0.8)}{\sqrt{(1-0.8)^2+(0-0.8)^2+(1-0.8)^2+(1-0.8)^2+(1-0.8)^2} \sqrt{(1-0.8)^2+(1-0.8)^2+(0-0.8)^2+(1-0.8)^2+(1-0.8)^2}} =$

$\frac{0.04 - 0.16 - 0.16 + 0.04 + 0.04}{\sqrt{0.04+0.64+0.04+0.04+0.04} \sqrt{0.04+0.04+0.64+0.04+0.04}} = \frac{-0.32 + 0.12}{\sqrt{0.64+0.16} \sqrt{0.64+0.16}} = \frac{-0.20}{\sqrt{0.8} \sqrt{0.8}} = \frac{-0.20}{0.8} = -0.25$

Jaccard similarity = $\frac{f_{11}}{f_{11}+f_{01}+f_{10}} = \frac{3}{3+1+1} = 0.6$

(b) x = (1,-2,2,-3), y = (-1,2,-3,-1)

cos(x,y) = $\frac{x.y}{||x|| . ||y||} = \frac{(1,-2,2,-3) * (-1,2,-3,-1)}{\sqrt{1^2+-2^2+2^2+-3^2} * \sqrt{-1^2+2^2+-3^2+-1^2}} = \frac{(-1-4-6+3)}{\sqrt{18} * \sqrt{15}} = \frac{-8}{\sqrt{18} * \sqrt{15}} =$

$\frac{-8}{3\sqrt{2} * \sqrt{15}} \simeq -0.49$

Euclidean distance, d(x,y) = $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} =$

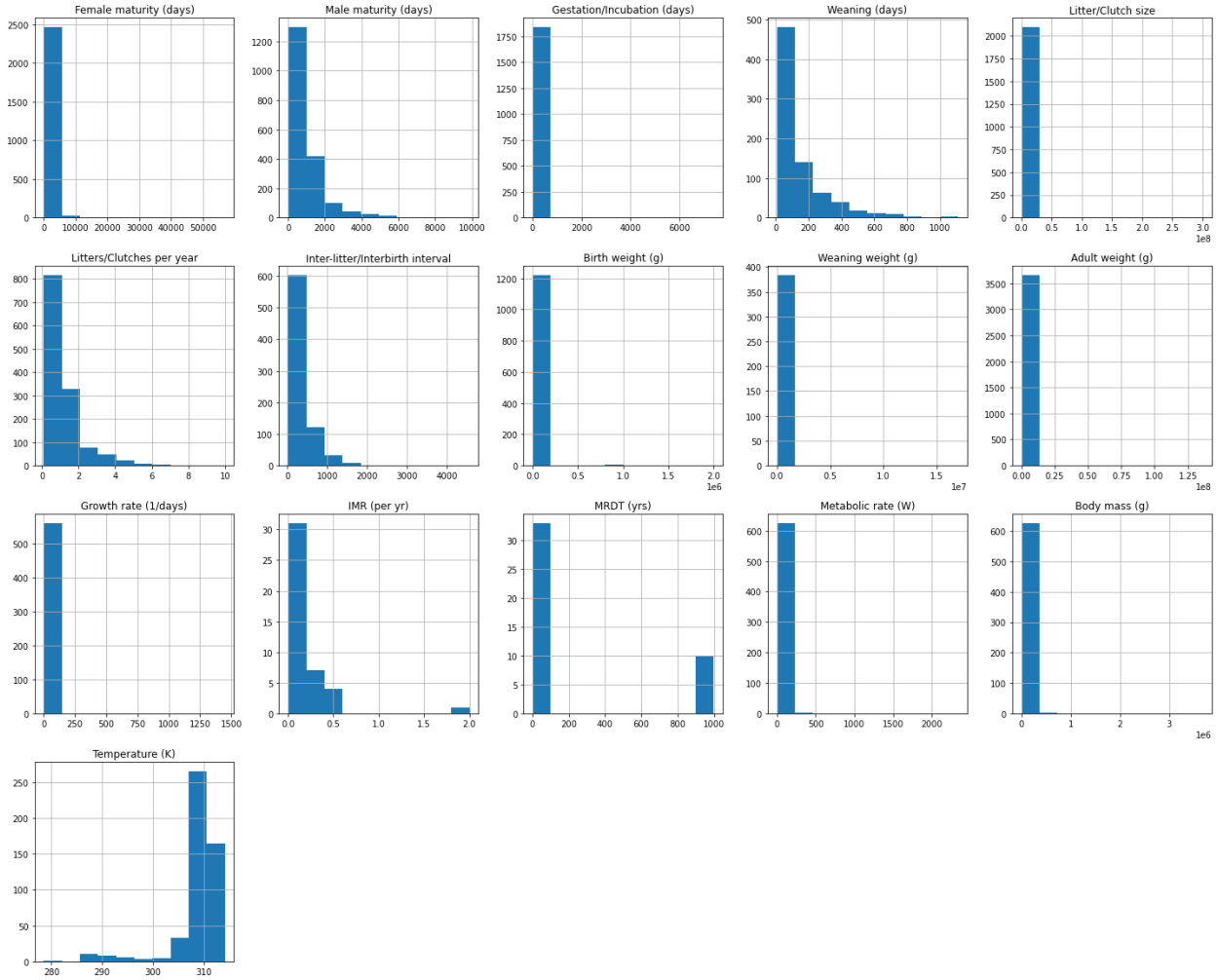
$\sqrt{(1+1)^2 + (-2-2)^2 + (2+3)^2 + (-3+1)^2} = \sqrt{4 + 16 + 25 + 4} = \sqrt{4 + 16 + 25 + 4} = 7$

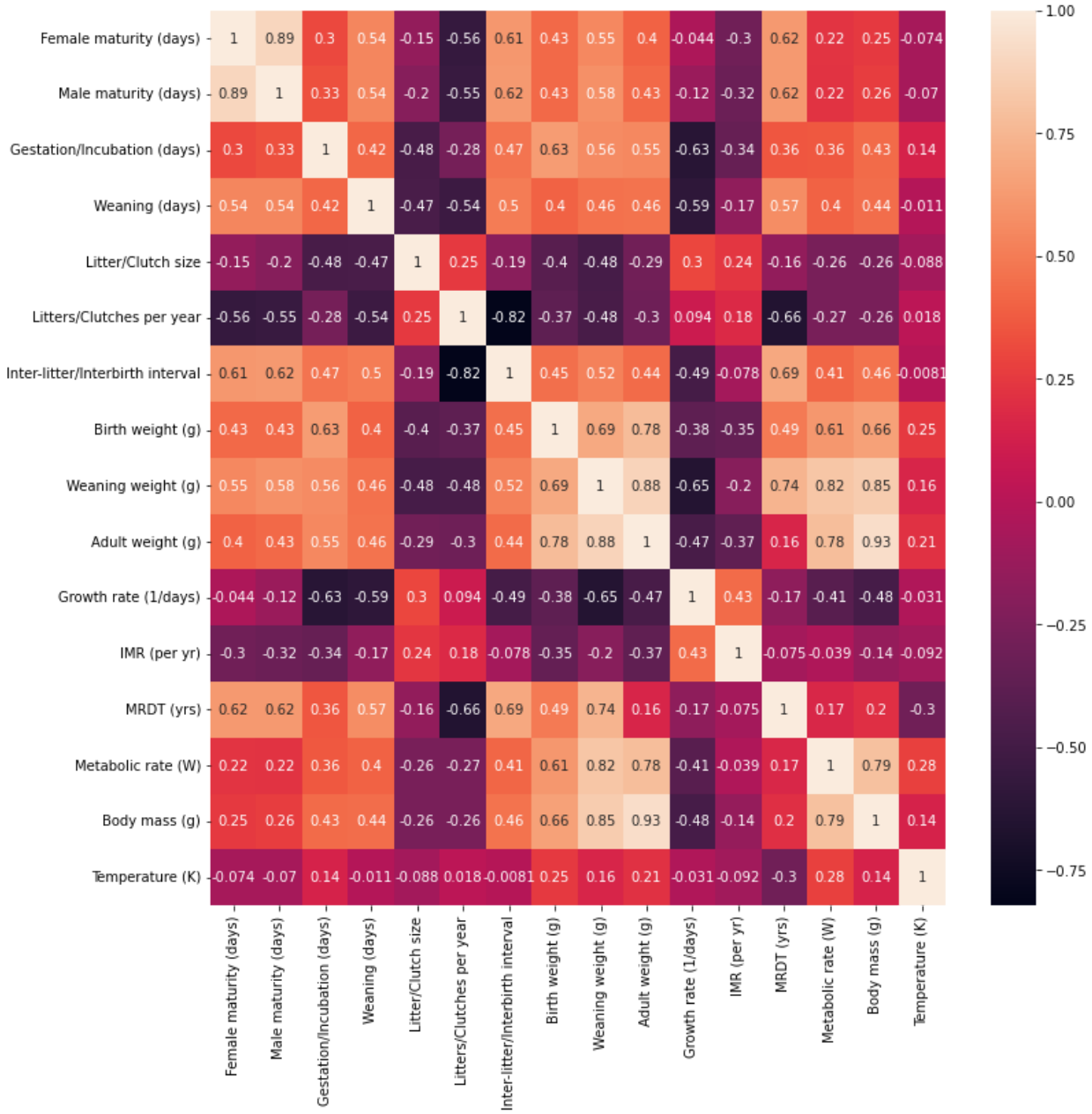
$$\text{corr}(x,y) = \frac{(1+0.5)(-1-2.2) + (-2+0.5)(2-2.2) + (2+0.5)(-3-2.2) + (-3+0.5)(-1-2.2)}{\sqrt{(1+0.5)^2+(-2+0.5)^2+(2+0.5)^2+(-3+0.5)^2}\sqrt{(-1-2.2)^2+(2-2.2)^2+(-3-2.2)^2+(-1-2.2)^2}} =$$

$$\frac{(1.5)(-3.2) + (-1.5)(-0.2) + (2.5)(-5.2) + (-2.5)(-3.2)}{\sqrt{(1.5)^2+(-1.5)^2+(2.5)^2+(-2.5)^2}\sqrt{(-3.2)^2+(-0.2)^2+(-5.2)^2+(-3.2)^2}} =$$

$$\frac{-4.8 + 0.3 - 13 + 8}{\sqrt{2.25+2.25+6.25+6.25}\sqrt{10.24+0.04+27.04+10.24}} = -0.645$$

4.



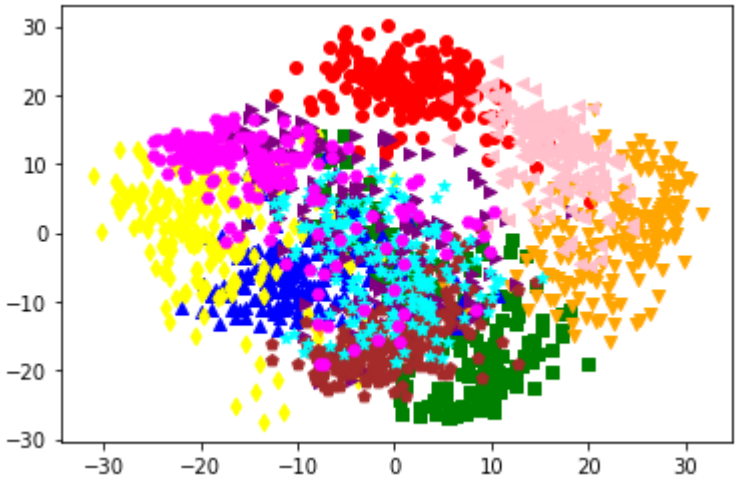


Hypothesis

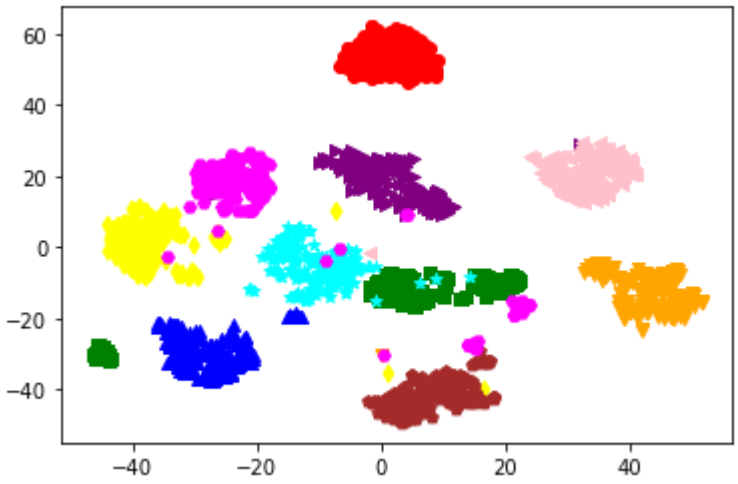
For each kingdom and class, since there are no null values present in each of the columns, we can qualify these as unique identifiers in our study. The hypothesis that can be tested from this dataset can be based on two variables that are strongly correlated namely: metabolic rate and body mass.

5.

PCA



t-SNE



6.