作业 1：数据认知与预处理

注意：要求给出必要的公式和中间计算过程，所有结果精确到小数点 2 位。

1. 现对数据集中的“年龄”收集到如下观测值：$32, 20, 30, 29, 18, 21, 24, 26$。

   a. 对以上属性分别进行均值为 0，方差为 1 的标准化（方差计算用有偏）。

   答：令均值为μ， 标准差为σ，则通过以下公式进行标准化：

$$x'_j = \frac{x_j - \mu}{\sigma}$$

根据题目，得到μ=25，σ=4.77(有偏)

代入以上公式，得到标准化后的观测值为：

$\boldsymbol{x'}$ =[1.47， -1.05，1.05，0.84， -1.47，-0.84，-0.21，0.21]

   b. 对以上属性进行[0,1]标准化，即最小值为 0， 最大值为 1。

   答：令 max 和 min 表示原观测值的最大和最小值，则通过以下公式得到标准化后的x'_j

$$x'_j = \frac{x_j - min}{max - min}$$

从题目可知 max=32, min=18 （如果题目给定最大最小值则用题目给定值），代入以上公式得到标准化后的观测值为：

$\boldsymbol{x'}$ =[1， 0.14，0.86，0.79， 0，0.21，0.43，0.57]

2. 现采集到以下蘑菇数据集，其中 NA 表示缺失值，请用每个类别的中位数/众数填补缺失值。假设颜色的可能取值为：{红色、褐色、白色、棕色}。

编号	尺寸	颜色	类别
1	2.3	红色	有毒
2	2.4	褐色	无毒
3	1.8	红色	有毒
4	NA	褐色	无毒
5	1.6	白色	无毒
6	2.4	NA	有毒
7	1.5	棕色	无毒

答：

表格中一共有两个缺失值，分别是标记为“无毒”的样本 4 的尺寸缺失和标记为“有毒“的样本 6 的颜色缺失。

属于“无毒”的样本 {2，5，7}在尺寸上有取值，按从小到大排序为：1.5， 1.6， 2.4，所以“无毒”类别对应尺寸的中位数是 1.6。根据题意，样本 4 的尺寸缺失值用 1.6 来填充。

属于“有毒”的样本{1，3}在颜色上有取值，均为红色，所以该类别颜色的众数是红色。根据题意，样本 6 的颜色用红色填充。

3. 给定$\boldsymbol{x}=[1,0,-1]^T, \boldsymbol{y}=[-1,1,0]^T, \boldsymbol{z}=[-2,1,-1]^T$，计算这三个向量两两之间的欧式距离、曼哈顿距离和余弦相似度。

答:

欧式距离: $d_{Euc}(\boldsymbol{x},\boldsymbol{y})=\sqrt{(1-(-1))^2+(0-1)^2+(-1-0)^2}=\sqrt{6}$=2.45

$$d_{Euc}(\boldsymbol{x},\boldsymbol{z})=\sqrt{(1-(-2))^2+(0-1)^2+(-1-(-1))^2}=\sqrt{10}=3.16$$

$$d_{Euc}(\boldsymbol{y},\boldsymbol{z})=\sqrt{(-1-(-2))^2+(1-1)^2+(0-(-1))^2}=\sqrt{2}=1.41$$

曼哈顿距离:

$$d_{Manh}(\boldsymbol{x},\boldsymbol{y})=|1-(-1)|+|0-1|+|-1-0|=4$$
$$d_{Manh}(\boldsymbol{x},\boldsymbol{z})=|1-(-2)|+|0-1|+|-1-(-1)|=4$$
$$d_{Manh}(\boldsymbol{y},\boldsymbol{z})=|-1-(-2)|+|1-1|+|0-(-1)|=2$$

余弦相似度:

$$s_{cos}(\boldsymbol{x},\boldsymbol{y})=\frac{x^Ty}{\|x\|\|y\|}=\frac{-1}{\sqrt{2}\sqrt{2}}=-\frac{1}{2}=-0.5$$

$$s_{cos}(\boldsymbol{x},\boldsymbol{z})=\frac{x^Tz}{\|x\|\|y\|}=\frac{-2+1}{\sqrt{2}\sqrt{6}}=-\frac{\sqrt{3}}{6}=-0.29$$

$$s_{cos}(\boldsymbol{y},\boldsymbol{z})=\frac{y^Tz}{\|y\|\|z\|}=\frac{2+1}{\sqrt{2}\sqrt{6}}=\frac{\sqrt{3}}{2}=0.87$$

4. 简答：(1) 哪些原因导致实际应用中收集到的数据往往存在噪声和缺失值？ (2) 归一化的主要作用是什么？

答 (1) 导致数据中存在噪声和缺失值的原因主要包括: 数据采集设备故障、数据传输、文件转换时发生的丢失、人为或计算机输入错误等。

(2) 不同特征的取值范围可能存在较大不同，归一化的主要作用是使所有特征取值服从同一分布或同一个取值范围，避免某些特征信息被忽略，提高算法建模时的有效性。