IEMS5726B
Data Science in Practice (Summer 2024) Assignment 2
JIANG Xijie, 1155202866

1.

a. Answer

1) Meaningless header in column 1-11.
2) Mullite value in column 11 To make these variable can be input of data analytics.
3) Missing value.
4) Inconsistent Name Format.
5) Outlier value like ‘77’.

b. Answer

1) Find out the meaning of the column or guess it.
2) Separate the value and one-hot encode create dummies variable.
3) Drop the observations with too many missing values.
4) Change all the Name with format of Family name then Given name like “Chen Bob”.
5) Drop the observations or using mean to fill.

c. Answer

Name	Age	1	2	3	4	5	6	7	8	9	10	Happy	Fear	Sad	Anger
CHAN Amy	22	6	6	6	6	3	4	6	6	6	6	1	1	0	0
Chen Bob	25	5	5	6	5	6	4	5	4	5	6	0	0	1	1
Zhen Chi	30	2	3	2	4	1	3	4	5	6		0	1	1	1

d. Answer

Tokenization, Term Weighting, Part-of-Speech Tagging.

e. Answer

Prevent Overfitting and creating variations of the audio data, which helps in training more generalizable models.