Data Manipulation and Data Wrangling

Data Manipulation is the process of changing data to make it easier to read or more organized.

Data Wrangling is process of transforming and mapping data from one raw data into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics

In [4]:
```
#importing all the important libraies
import cv2
import numpy as np
import pandas as pd
import os
import glob
```

In [5]:
```
# importing metadata from the directory
df = pd.read_csv("HAM10000_metadata.csv")
df.head(5)
```

Out[5]:

	lesion_id	image_id	dx	dx_type	age	sex	localization
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear

In [6]:
```
# to show the categories of lesion
np.unique(df['dx'].tolist())
```

Out[6]:
```
array(['akiec', 'bcc', 'bkl', 'df', 'mel', 'nv', 'vasc'], dtype='<U5')
```

In [7]:
```
def image_to_feature_vector(image, size=(32, 32)):
    # resize the image to a fixed size, then flatten the image into
    # a list of raw pixel intensities
    return cv2.resize(image, size).flatten()
```

In [8]:
```
img_dir = "C:\\Users\\DELL\\3D Objects\\skin-cancer-mnist-ham10000\\HAM10000_images_part_1"
# Enter Directory of all images
data_path = os.path.join(img_dir,'*g')
files = glob.glob(data_path)
data=[]
F1=[]

for f1 in files:
    word_list= f1.split('\\')                          # spliting the path of each file with
"//"
    F1.append(word_list[-1].split('.')[0])             # spliting the image Id (eg.ISIC_0034
320.jpg) of each file with "."
    img = cv2.imread(f1)
    images=image_to_feature_vector(img, size=(32, 32)) # resizeing of an image
    images1=images.tolist()
    data.append(images1)
```

In [9]:
```
# number of images in the data
len(data)
```

Out[9]:
```
10015
```

In [84]:
```
# number of pixels in the data
len(data[0])
```

Out[84]:
```
3072
```

In [91]:
```
a=['image_id']
str1='Pixel_'
for i in range(3072):
   a.append(str1 + str(i))
```

In [66]:
```
# Python3 program to Convert 1D
# list to 2D list
from itertools import islice

def convert(lst, var_lst):
    it = iter(lst)
    F2=[list(islice(it, i)) for i in var_lst]
    return F2

# Driver code
var_lst = [1]*len(F1)
F2=convert(F1, var_lst)
```

In [93]:
```
def merge(lst1, lst2):
    return [a + b for (a, b) in zip(lst1, lst2)]
F3=merge(F2,data)
```

In [94]:
```
df1=pd.DataFrame(F3,columns=a)
df1.head(5)
```

In [114]:
```
# labeling the categories of lesion
def score_to_numeric(x):
    if x=='akiec':
        return 0
    if x=='bcc':
        return 1
    if x=='bkl':
        return 2
    if x=='df':
        return 3
    if x=="mel":
        return 4
    if x=='nv':
        return 5
    if x=='vasc':
        return 6
```

In [116]:
```
#df=df.drop('label',axis=1)
```

In [119]:
```
# merging the metadata with pixels of the images according to the image_ID
df3=pd.merge(df, df1, on='image_id', how='outer')
df3.head(5)
```

Out[119]:

	lesion_id	image_id	dx	dx_type	age	sex	localization	Pixel_0	Pixel_1	Pixel_2	...	Pixel_3062	Pixel_3063	P
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp	187	148	190	...	178	154	
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp	25	14	23	...	91	43	
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp	146	133	186	...	167	143	
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp	27	16	31	...	77	22	
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear	134	110	153	...	219	179	

5 rows × 3079 columns

In [121]:
```
df3['label']=df['dx'].apply(score_to_numeric)
df3.head(5)
```

Out[121]:

	lesion_id	image_id	dx	dx_type	age	sex	localization	Pixel_0	Pixel_1	Pixel_2	...	Pixel_3063	Pixel_3064	P
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp	187	148	190	...	154	132	
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp	25	14	23	...	43	26	
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp	146	133	186	...	143	128	
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp	27	16	31	...	22	16	
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear	134	110	153	...	179	161	

5 rows × 3080 columns

In [ ]:
```
# saving the dataframe
df3.to_csv('df3_final.csv', header=True, index=False)
```