DataLab Nigeria Data Science Virtual Class

PANDAS COURSE

Telstra is a telecommunication company on a journey to enhance the customer experience - ensuring everyone in the company is putting customers first. Using dataset of features from their service logs, you are tasked with merging these datasets in preparation for future analysis.

Find the datasets (files) attached to the accompanying email and how your output should look like after the merging.

The following operations should also proceed your merging

1. Import the neccessary libraries for this task:

import numpy as np

import pandas as pd

2. Read in all the files to be merge with "pd.read_csv".

3. Check for the data type of each of the files using ".dtypes":

E.g for event_type file:

event_type.dtypes

4. If the **id column** in any file is found as "object or character", convert it to numeric data type. Use this code snippet: E.g To perform for the "event_type" file use:

event_type['id']=pd.to_numeric(event_type['id'],errors='coerce')

5. Now merge the entire files (5 of them). Check the dimensions and get the amount of missing data arising from your joining operations.

6. Finally do series of sub-setting on your final joined dataset by attempting to explore the final joined dataset by selecting columns of your choice.

FIND FURTHER CODE SNIPPETS BELOW TO ASSIST YOU IN THE TASK

```
#merging the data sets to have all the available info

train_1 = train.merge(severity_type, how = 'left', left_on='id', right_on='id')
train_2 = train_1.merge(resource_type, how = 'left', left_on='id', right_on='id')
train_3 = train_2.merge(log_failure, how = 'left', left_on='id', right_on='id')
train_4 = train_3.merge(event_type, how = 'left', left_on='id', right_on='id')
```

```
#dropping the duplicate records

train_4.drop_duplicates(subset= 'id', keep= 'first', inplace = True)
```

```
#count plot for fault severity

plt.figure(figsize = (8,6))
sns.countplot(train_4['fault_severity'])
plt.show()

#count plot for severity type

plt.figure(figsize = (8,6))
sns.countplot(train_4['severity_type'])
plt.show()
```