NIT HACKATHON

Vertical : Data Science

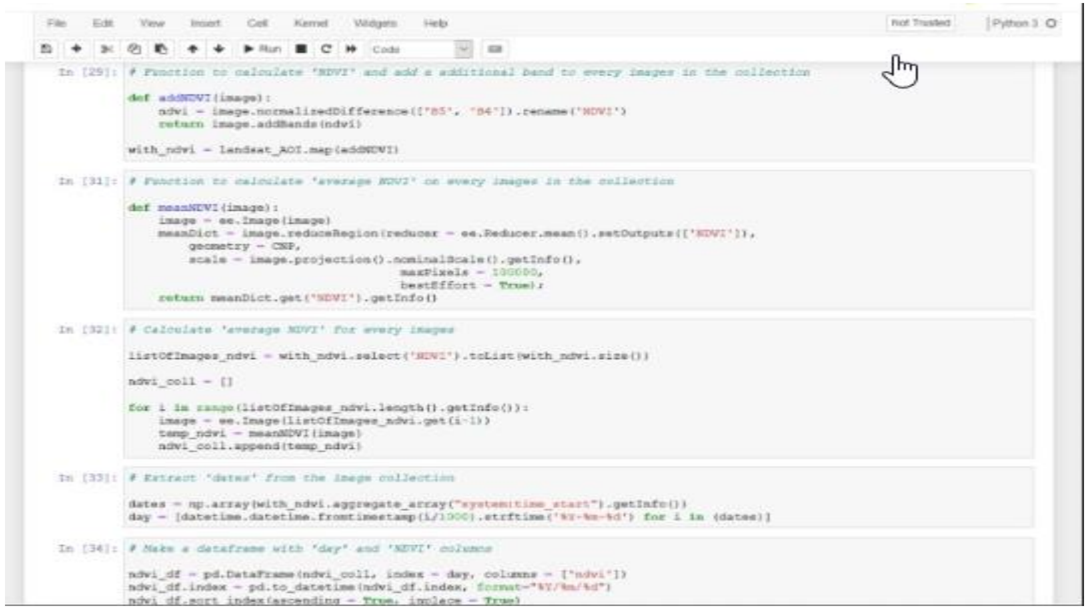
Topic : Website called ‘Pipeline’ for data preprocessing

College : Shiv Nadar University Chennai

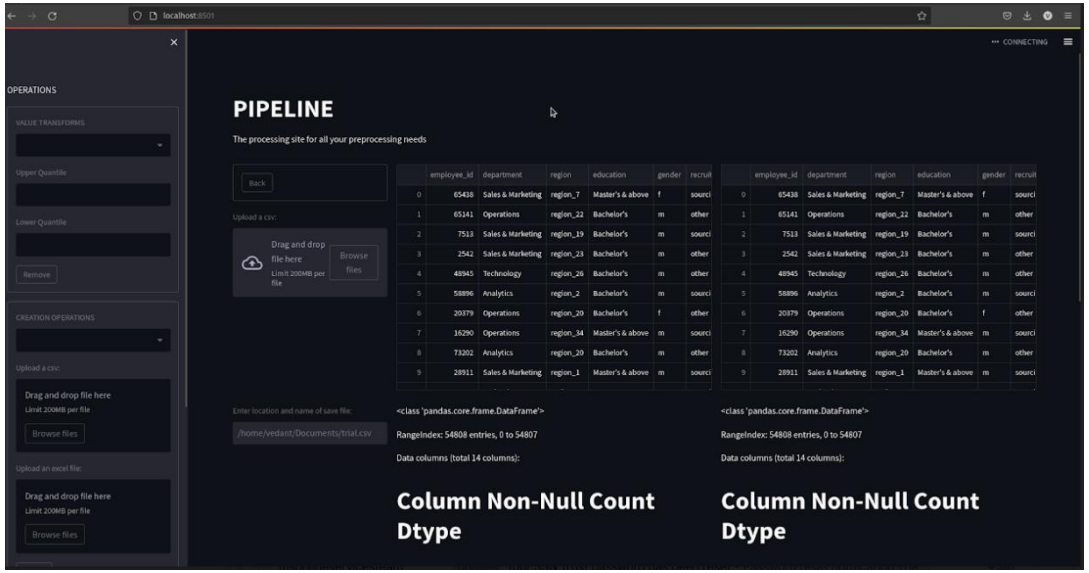
Team : Rhea Pandita (B.TECH 2nd Year)
Vedant Nair (B.TECH 2nd Year)

Aim of the project

- ‘Pipeline’ for Data Preprocessing is a website meant to aid in the data preprocessing of csv files.
- The usual process of preprocessing csv files includes writing 10-20 lines of code. Also, the methods applied differ from one file to another based on its contents.
- This website’s aim is to reduce the time taken to do the same by usage of drop-down menus back-ended by common methods data analysts apply.



```
In [29]: # Function to calculate 'NDVI' and add a additional band to every images in the collection

         def addNDVI(image):
             ndvi = image.normalizedDifference(['B5', 'B4']).rename('NDVI')
             return image.addBands(ndvi)

         with_ndvi = landsat_AOI.map(addNDVI)

In [31]: # Function to calculate 'average NDVI' on every images in the collection

         def meanNDVI(image):
             image = ee.Image(image)
             meanDict = image.reduceRegion(reducer = ee.Reducer.mean().setOutputs(['NDVI']),
                 geometry = CNP,
                 scale = image.projection().nominalScale().getInfo(),
                                             maxPixels = 100000,
                                             bestEffort = True);
             return meanDict.get('NDVI').getInfo()

In [32]: # Calculate 'average NDVI' for every images

         listOfImages_ndvi = with_ndvi.select('NDVI').toList(with_ndvi.size())

         ndvi_coll = []

         for i in range(listOfImages_ndvi.length().getInfo()):
             image = ee.Image(listOfImages_ndvi.get(i-1))
             temp_ndvi = meanNDVI(image)
             ndvi_coll.append(temp_ndvi)

In [33]: # Extract 'dates' from the image collection

         dates = np.array(with_ndvi.aggregate_array("system:time_start").getInfo())
         day = [datetime.datetime.fromtimestamp(i/1000).strftime('%Y-%m-%d') for i in (dates)]

In [34]: # Make a dataframe with 'day' and 'NDVI' columns

         ndvi_df = pd.DataFrame(ndvi_coll, index = day, columns = ["ndvi"])
         ndvi_df.index = pd.to_datetime(ndvi_df.index, format="%Y/%m/%d")
         ndvi_df.sort_index(ascending = True, inplace = True)
```



PIPELINE

The processing site for all your preprocessing needs

Enter location and name of save file: /home/vedant/Documents/trial.csv

	employee_id	department	region	education	gender	recruit
0	65438	Sales & Marketing	region_7	Master's & above	f	sourci
1	65141	Operations	region_22	Bachelor's	m	other
2	7513	Sales & Marketing	region_19	Bachelor's	m	sourci
3	2542	Sales & Marketing	region_23	Bachelor's	m	other
4	48945	Technology	region_26	Bachelor's	m	other
5	58896	Analytics	region_2	Bachelor's	m	sourci
6	20379	Operations	region_20	Bachelor's	f	other
7	16290	Operations	region_34	Master's & above	m	sourci
8	73202	Analytics	region_20	Bachelor's	m	other
9	28911	Sales & Marketing	region_1	Master's & above	m	sourci

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 54808 entries, 0 to 54807

Data columns (total 14 columns):

Column Non-Null Count Dtype

	employee_id	department	region	education	gender	recruit
0	65438	Sales & Marketing	region_7	Master's & above	f	sourci
1	65141	Operations	region_22	Bachelor's	m	other
2	7513	Sales & Marketing	region_19	Bachelor's	m	sourci
3	2542	Sales & Marketing	region_23	Bachelor's	m	other
4	48945	Technology	region_26	Bachelor's	m	other
5	58896	Analytics	region_2	Bachelor's	m	sourci
6	20379	Operations	region_20	Bachelor's	f	other
7	16290	Operations	region_34	Master's & above	m	sourci
8	73202	Analytics	region_20	Bachelor's	m	other
9	28911	Sales & Marketing	region_1	Master's & above	m	sourci

<class 'pandas.core.frame.DataFrame'>

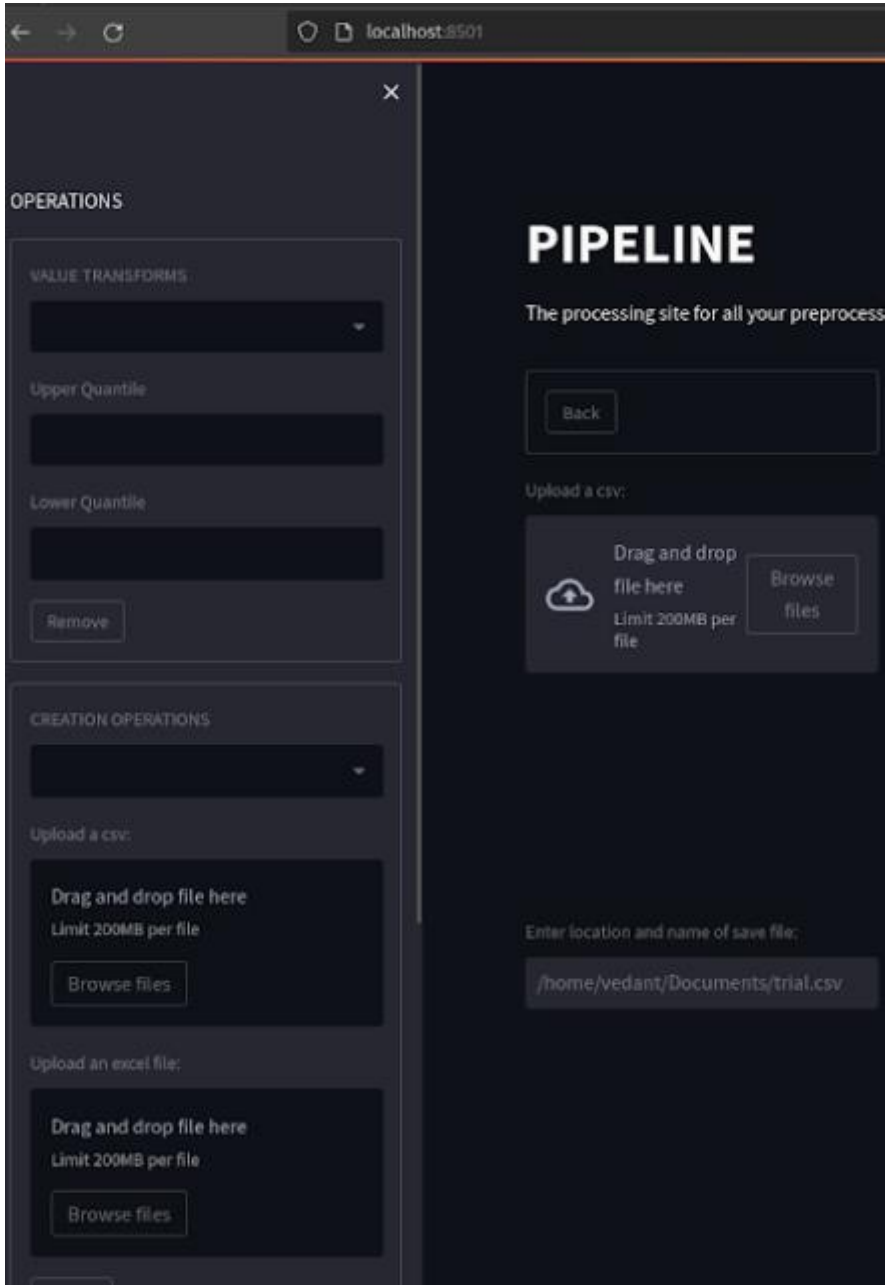
RangeIndex: 54808 entries, 0 to 54807

Data columns (total 14 columns):

Column Non-Null Count Dtype

Features available

- Uploading required file from the system or using drag and drop
- Dropdown menus to choose data preprocessing method from
- Following are the preprocessing dropdowns:
  - Null Value Treatment
  - Outlier Treatment
  - Dropping Data
  - Data Creation
  - Feature Scaling



localhost:8501

OPERATIONS

VALUE TRANSFORMS

Upper Quantile

Lower Quantile

CREATION OPERATIONS

Upload a csv:

Drag and drop file here
Limit 200MB per file

Upload an excel file:

Drag and drop file here
Limit 200MB per file

PIPELINE

The processing site for all your preprocess

Upload a csv:

Drag and drop file here
Limit 200MB per file

Enter location and name of save file:

/home/vedant/Documents/trial.csv

Scope of the project

- This project can be used to preprocess data that is in table form. Since a large number of industries are required to have their own repository of their business in the past and the present, this will facilitate a quick cleanup of the data

- Since most industries are rapidly moving forward in the direction of online enterprise, it is necessary for them to tabulate all changes that are encountered (even the minor ones). This increases the number of databases overall and its size, thus demanding a need for cleanup. This can greatly benefit the medical industry, the judicial system keep track of their cases etc and many more places.