PYTHON FOR FEEDBACK DATA CLEANING AND DATA ANALYSIS

- Import Panda File And Read train.csv File

```
import pandas as pd
Review=pd.read_csv("C:/Users/ARCHIES/Documents/Power BI/train.csv")
```

- Pandas .shape is used to return shape of data frames and series.

```
Review.shape
```

```
(22551, 7)
```

- Pandas head() method is used to return top n (5 by default) rows of a data frame or series.

```
Review.head()
```

	id	host_id	Feedback	Is_Response	User_ID	Browser_Used	Device_Used
0	2015	2217	The room was kind of clean but had a VERY stro...	not happy	id10326	Edge	Mobile
1	2695	2986	I stayed at the Crown Plaza April -- - April -...	happy	id10327	Internet Explorer	Mobile
2	3176	3718	I booked this hotel through Hotwire at the low...	happy	id10328	Mozilla	Tablet
3	3309	4108	Stayed here with husband and sons on the way t...	not happy	id10329	InternetExplorer	Desktop
4	7071	17391	My girlfriends and I stayed here to celebrate ...	happy	id10330	Edge	Tablet

- Calculating Percentage and Count of Missing Values

```
count = Review.isnull().sum().sort_values(ascending=False)
percentage = ((Review.isnull().sum()/len(Review)*100)).sort_values(ascending=False)
missing_data = pd.concat([count,percentage], axis=1,
keys=['Count','Percentage'])

print('Count and Percentage of Missing Values: ')
missing_data
```

```
Count and Percentage of Missing Values:
```

	Count	Percentage
id	0	0.0
host_id	0	0.0
Feedback	0	0.0
Is_Response	0	0.0
User_ID	0	0.0
Browser_Used	0	0.0
Device_Used	0	0.0

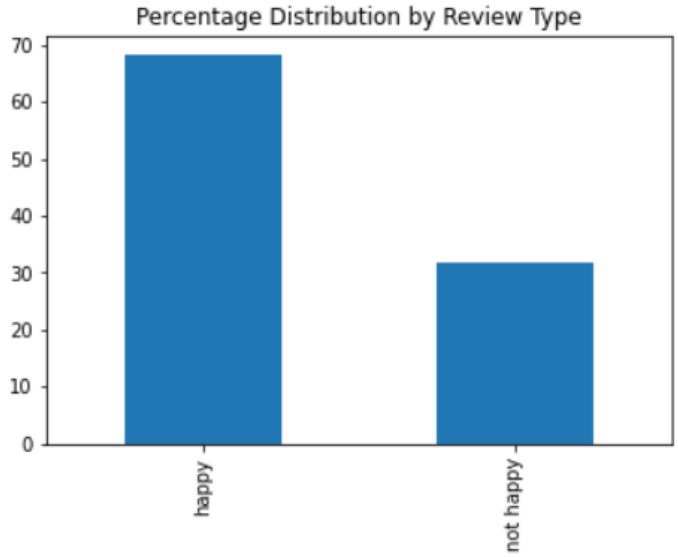
- Matplotlib is also called magic functions. It draw inline plots for quick data analysis. By using Matplotlib it representing Percentage Distribution by Review Type i.e Happy & Not Happy.

```
import matplotlib.pyplot as plt
%matplotlib inline
print('Percentage for default\n')
print(round(Review.Is_Response.value_counts(normalize=True)*100,2))
round(Review.Is_Response.value_counts(normalize=True)*100,2).plot(kind='bar')
plt.title('Percentage Distribution by Review Type')
plt.show()
```

```
Matplotlib is building the font cache; this may take a moment.
```

```
Percentage for default

happy        68.19
not happy    31.81
Name: Is_Response, dtype: float64
```



- Importing Regular Expression & String module. Define Lambda Expression Named Cleaned_Data.

```
import re
import string

def Data_Cleaning (text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '',text)
    text = re.sub('\w*\d\w*', '',text)
    return text

Cleaned_Data = lambda x: Data_Cleaning(x)
```

- Feedback of Cleaned_Feedback with the help of head().

```
import pandas as pd
Review=pd.read_csv("C:/Users/praja/OneDrive/Documents/train.csv")
Review['Cleaned_Feedback'] = pd.DataFrame(Review.Feedback.apply(Cleaned_Data))
Review.head()
```

	id	host_id	Feedback	Is_Response	User_ID	Browser_Used	Device_Used	Cleaned_Feedback
0	2015	2217	The room was kind of clean but had a VERY stro...	not happy	id10326	Edge	Mobile	the room was kind of clean but had a very stro...
1	2695	2986	I stayed at the Crown Plaza April -- - April -...	happy	id10327	Internet Explorer	Mobile	i stayed at the crown plaza april april th...
2	3176	3718	I booked this hotel through Hotwire at the low...	happy	id10328	Mozilla	Tablet	i booked this hotel through hotwire at the low...
3	3309	4108	Stayed here with husband and sons on the way t...	not happy	id10329	InternetExplorer	Desktop	stayed here with husband and sons on the way t...
4	7071	17391	My girlfriends and I stayed here to celebrate ...	happy	id10330	Edge	Tablet	my girlfriends and i stayed here to celebrate ...

- Define a Lambda Expression named Cleaned_Data2.

```
def Data_Cleaning2 (text):
    text = re.sub('[''""...]', '', text)
    text = re.sub('\n','', text)
    return text

Cleaned_Data2 = lambda x: Data_Cleaning2(x)
```

- Feedback of Cleaned_Feedback2 with the help of head().

```
Review['Cleaned_Feedback2'] = pd.DataFrame(Review['Cleaned_Feedback'].apply(Cleaned_Data2))
Review.head()
```

	id	host_id	Feedback	Is_Response	User_ID	Browser_Used	Device_Used	Cleaned_Feedback	Cleaned_Feedback2
0	2015	2217	The room was kind of clean but had a VERY stro...	not happy	id10326	Edge	Mobile	the room was kind of clean but had a very stro...	the room was kind of clean but had a very stro...
1	2695	2986	I stayed at the Crown Plaza April -- - April -...	happy	id10327	Internet Explorer	Mobile	i stayed at the crown plaza april april th...	i stayed at the crown plaza april april th...
2	3176	3718	I booked this hotel through Hotwire at the low...	happy	id10328	Mozilla	Tablet	i booked this hotel through hotwire at the low...	i booked this hotel through hotwire at the low...
3	3309	4108	Stayed here with husband and sons on the way t...	not happy	id10329	InternetExplorer	Desktop	stayed here with husband and sons on the way t...	stayed here with husband and sons on the way t...
4	7071	17391	My girlfriends and I stayed here to celebrate ...	happy	id10330	Edge	Tablet	my girlfriends and i stayed here to celebrate ...	my girlfriends and i stayed here to celebrate ...

- Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

```
from sklearn.feature_extraction.text import TfidVectorizer
from sklearn.linear_model import LogisticRegression

tvec = TfidVectorizer()
clf2 = LogisticRegression(solver = 'lbfgs')

from sklearn.pipeline import Pipeline
```