*This is data pre processing *

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Importing the data as a dataframe and storing into a variable called data

```
data = pd.read_csv(r'airlines.csv')
```

Giving name to the columns present in the dataset

```
data.columns = ['Airline ID', 'Name', 'Alias', 'IATA', 'ICAO', 'Callsign', 'Country',
```

The first 5 entries of the dataset

```
print(data.head())
```

```
   Airline ID  ... Active
0          -1  ...      Y
1           1  ...      Y
2           2  ...      N
3           3  ...      Y
4           4  ...      N

[5 rows x 8 columns]
```

The shape of the data

```
print(data.shape)
```

```
(6162, 8)
```

Changing \N value to null which is easily recognised by inbuilt functions

```
data['Alias'] = data['Alias'].replace(r'\N', np.NaN)
data['Callsign'] = data['Callsign'].replace(r'\N', np.NaN)
data['ICAO'].replace(r'\N', np.NaN, inplace = True)
data['Active'].replace(r'n', 'N', inplace = True)
```

Counting the number of missing values in each attributes of the dataset

```
print(data.isnull().sum())
```

```
Airline ID       0
Name             0
Alias         5984
IATA          4627
ICAO           275
Callsign       811
Country         15
Active           0
dtype: int64
```

Dropping these as their missing values are greater than 50%

```
data.drop(columns = ['IATA', 'Alias'], axis = 1 ,inplace = True)
```

Filling the missing value of countires with its mode

```
data['Country'].fillna(data['Country'].mode()[0], inplace = True)
```

These two attributes are supposed to have unique values, so dropping all the rows that hve the val

```
data.dropna(subset=['Callsign', 'ICAO'], inplace=True)
```

Counting the number of missing values in each attributes of the dataset

```
print(data.isnull().sum())
```

```
Airline ID    0
Name          0
ICAO          0
Callsign      0
Country       0
Active        0
dtype: int64
```

Final shape of the clean dataset

```
print(data.shape)
```

```
(5297, 6)
```

Saving it into a csv

```
    data.to_csv('cleaned_data.csv', index = False)
```

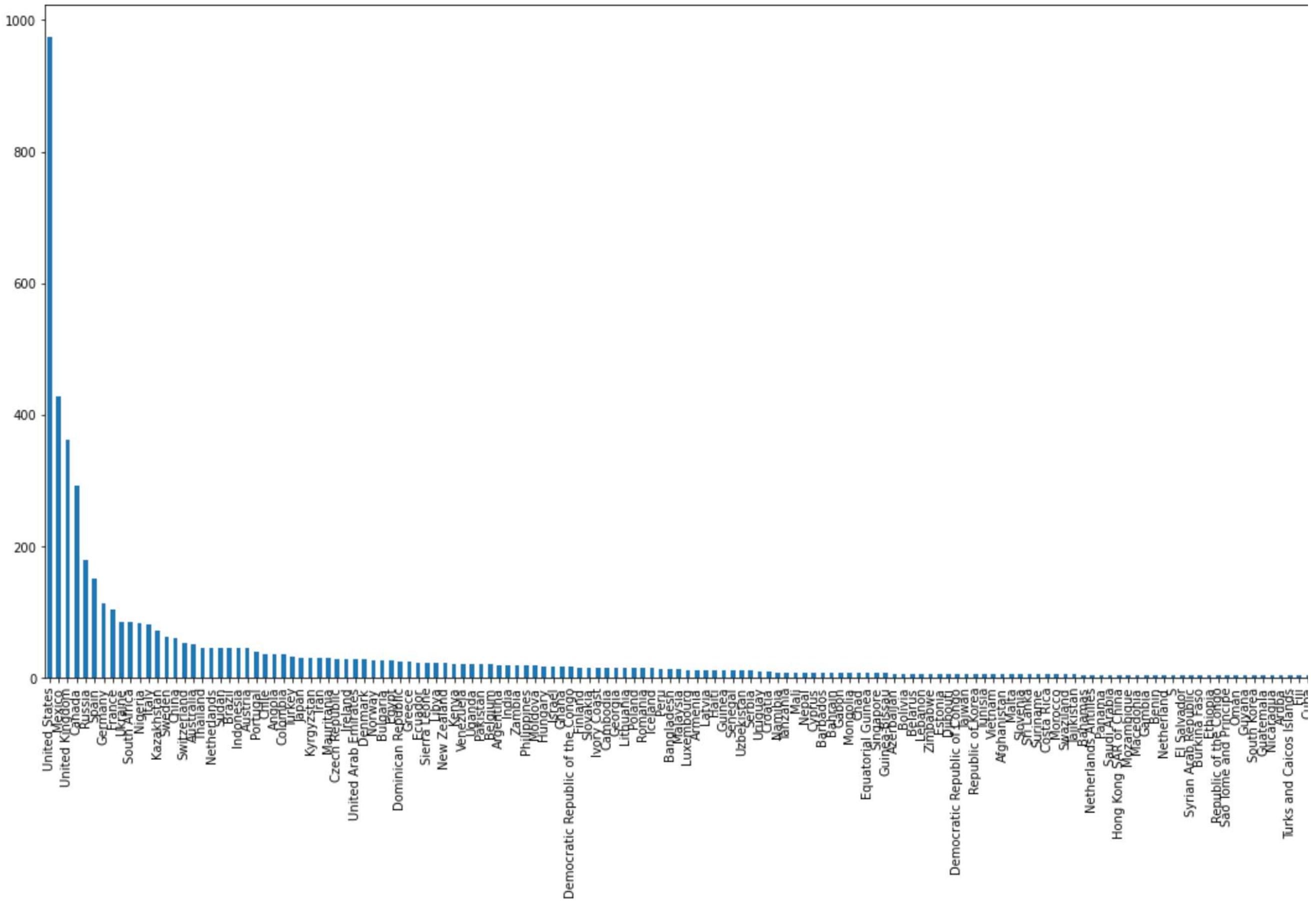
Visualising the distribution of categorical variable, and their unique values

```
for column in data.columns:
        if(column in ['Airline ID', 'Name']):
            pass
        elif(column in ['ICAO', 'Callsign']):
            print(column)
            print("The unique values of this column is : ")
            print(data[column].unique())
        else:
            print(column)
            print("The unique values of this column is : ")
            print(data[column].unique())
            if(column == 'Active'):
                data[column].value_counts().plot(kind = 'pie')
            else:
                data[column].value_counts().plot(kind = 'bar', figsize=(35, 10))
            plt.show()
        print("\n\n")
```

```
ICAO
The unique values of this column is :
['GNL' 'RNX' 'CHD' ... 'CBG' 'SSX' 'SJM']
```

```
Callsign
The unique values of this column is :
['GENERAL' 'NEXTIME' 'CHKALOVSK-AVIA' ... 'SPRAY' 'Shasta'
 'RussianConnecty']
```

```
Country
The unique values of this column is :
['United States' 'South Africa' 'Russia' 'United Kingdom' 'Thailand'
 'Canada' 'Australia' 'Belgium' 'Mexico' 'Spain' 'France'
 'United Arab Emirates' 'Republic of Korea' 'Pakistan' 'Libya'
 'Ivory Coast' 'Ukraine' 'Democratic Republic of the Congo' 'Iran'
 'Finland' 'Brazil' 'Colombia' 'AEROCENTER' 'Ghana' 'Kenya' 'Togo'
 'Somali Republic' 'Morocco' 'Dominican Republic' 'Albania' 'Nigeria'
 'Germany' 'Slovenia' 'Czech Republic' 'Benin' 'AEROCESAR' 'Greece'
 'Chile' 'Tanzania' 'Bolivia' 'Italy' 'Sweden' 'Argentina' 'Sierra Leone'
 'Indonesia' 'Senegal' 'Afghanistan' 'Uganda' 'Bosnia and Herzegovina'
 'Gabon' 'Angola' 'Uzbekistan' 'Namibia' 'Turkey' 'Vietnam' 'Zambia'
 'Egypt' 'Ireland' 'Switzerland' 'Gambia' 'Serbia' 'Peru' 'Slovakia'
 'Denmark' 'Azerbaijan' 'AIRPORT HELICOPTER' 'Hong Kong' 'Croatia'
 'Hungary' 'Estonia' 'Swaziland' 'India' 'Reunion' 'Iceland' 'Israel'
 'Austria' 'Jamaica' 'Malta' 'Portugal' 'Japan' 'Cyprus' 'Kazakhstan'
 'Kyrgyzstan' 'Turkmenistan' 'Cambodia' 'Netherlands Antilles'
 'Sao Tome and Principe' 'Venezuela' 'ALNACIONAL' 'Lithuania' 'Maldives'
 'Malawi' 'Moldova' 'Montenegro' 'Macao' 'Seychelles' 'Bulgaria'
 'Papua New Guinea' 'Latvia' 'ANTARES' 'AVINOR' 'New Zealand' 'ALCON'
 'ASTORIA' 'COMPANY AS' 'Philippines' 'AIRPAC' 'AIRFLIGHT' 'ASA PESADA'
 'AIR PRINT' 'APPALACHIAN' 'ALASKA PACIFIC' 'AEROPERLAS' 'AEROPUMA' 'ATCO'
 'AVIOQUINTANA' 'Aruba' ' S.A.' 'ARMSTRONG' 'Armenia' 'AIREX' 'Chad'
 'ALASKA' 'SCHEFF' 'AEROSUN' 'ALL STAR' 'AIR CLASS' 'Nepal' 'Sudan'
 'Panama' 'Guinea-Bissau' 'ATLANTIS CANADA' 'Burkina Faso' 'Netherlands'
 'Ecuador' 'AUDI AIR' 'Uruguay' 'AURORA AIR' 'AUSA' 'AVIANCA' 'ALAMO'
 'Djibouti' 'AVEMEX' 'Vanuatu' 'AQUILINE' 'ACTIVE AERO' 'Bangladesh'
 'Georgia' 'El Salvador' 'AEROVARADERO' 'AIRNAT' 'Niger' 'Jordan'
 'AIRWAVE' 'AEROWEE' 'AIR FREIGHTER' 'Malaysia' 'RENTAXEL' 'Macedonia'
 'AIRMAN' 'ATLANTIC NICARAGUA' 'AZIMUT' 'Zimbabwe' 'ARIZAIR' 'Guatemala'
 'Bahrain' 'Barbados' 'Botswana' 'French Polynesia' 'Russian Federation'
 'China' 'Belize' 'Mozambique' 'Marshall Islands' 'Algeria' 'Ethiopia'
 'Air S' 'Fiji' 'Mali' 'Faroe Islands' 'Cameroon' 'Guinea' 'Belarus'
 'French Guiana' 'Haiti' 'Comoros' 'Honduras' 'Myanmar'
 "Democratic People's Republic of Korea" 'Mauritania' 'Mauritius'
 'Madagascar' 'Mongolia' 'Norway' 'Burundi' 'Sri Lanka' 'Romania'
 'Republic of the Congo' 'Nicaragua' 'Turks and Caicos Islands' 'Kiribati'
 'S' 'Bahamas' 'Suriname' 'Syrian Arab Republic' 'AIR-MAUR' 'Cape Verde'
 'Luxembourg' 'Oman' 'Antigua and Barbuda' 'Trinidad and Tobago'
 'Hong Kong SAR of China' 'Cayman Islands' 'Central African Republic'
 'Democratic Republic of Congo' 'Poland' 'Taiwan' 'ASUR' 'Cuba' 'DRAGON'
 'Bhutan' 'Equatorial Guinea' 'UNIFORM OSCAR' 'Russia]]' 'Eritrea'
 'Tunisia' 'Lebanon' 'Costa Rica' 'Saint Lucia' 'Monaco' 'Solomon Islands'
 'Liberia' 'ALDAWLYH AIR' 'Iraq' 'Kuwait' 'Singapore' 'ODINN' 'LAP'
 'Lao Peoples Democratic Republic' 'Bermuda' 'WATCHDOG' 'Montserrat'
 'Nauru' 'Saudi Arabia' 'Palau' 'Tonga' 'Samoa' 'Qatar' 'UNited Kingdom'
 'Netherland' 'Guyana' 'Brunei' 'Rwanda' 'ACOM' ' Boonville Stage Line'
 'Saint Vincent and the Grenadines' 'SWISSBIRD' 'Tajikistan' 'AEROSOL'
 'Saint Kitts and Nevis' 'Paraguay' 'VELES' "Cote d'Ivoire" 'Yemen' '\\N'
 'South Korea' 'Burma' 'Puerto Rico' 'Congo (Kinshasa)']
```



```
Active
The unique values of this column is :
['N' 'Y']
```