Importing Libraries

In [1]:
```python
import re
import warnings
import numpy as np
import pandas as pd
import seaborn as sns
from sklearn import metrics
import matplotlib.pyplot as plt
warnings.filterwarnings('ignore')
from matplotlib.gridspec import GridSpec
from sklearn.metrics import accuracy_score
from pandas.plotting import scatter_matrix
from sklearn.naive_bayes import MultinomialNB
from sklearn.multiclass import OneVsRestClassifier
from sklearn.neighbors import KNeighborsClassifier
```

Importing the Dataset

In [2]:
```python
cv = pd.read_csv("C:/Users/naikc/OneDrive/Desktop/Resume Analysis/UpdatedResumeDataSet.csv")
```

Exploratory Data Analysis

In [3]:
```python
cv.head(10)
```

Out[3]:

	Category	Resume
0	Data Science	Skills * Programming Languages: Python (pandas...
1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E...
2	Data Science	Areas of Interest Deep Learning, Control Syste...
3	Data Science	Skills â¢ R â¢ Python â¢ SAP HANA â¢ Table...
4	Data Science	Education Details \r\n MCA YMCAUST, Faridab...
5	Data Science	SKILLS C Basics, IOT, Python, MATLAB, Data Sci...
6	Data Science	Skills â¢ Python â¢ Tableau â¢ Data Visuali...
7	Data Science	Education Details \r\n B.Tech Rayat and Bahr...
8	Data Science	Personal Skills â ¢ Ability to quickly grasp t...
9	Data Science	Expertise â Data and Quantitative Analysis â...

In [4]:
```python
cv.tail(10)
```

Out[4]:

	Category	Resume
952	Testing	PERSONAL SKILLS â¢ Quick learner, â¢ Eagerne...
953	Testing	COMPUTER SKILLS & SOFTWARE KNOWLEDGE MS-Power ...
954	Testing	Skill Set OS Windows XP/7/8/8.1/10 Database MY...
955	Testing	â¢ Good logical and analytical skills â¢ Pos...
956	Testing	COMPUTER PROFICIENCY â¢ Basic: MS-Office (Pow...
957	Testing	Computer Skills: â¢ Proficient in MS office (...
958	Testing	â Willingness to accept the challenges. â ...
959	Testing	PERSONAL SKILLS â¢ Quick learner, â¢ Eagerne...
960	Testing	COMPUTER SKILLS & SOFTWARE KNOWLEDGE MS-Power ...
961	Testing	Skill Set OS Windows XP/7/8/8.1/10 Database MY...

In [5]:
```python
print('Count of columns in the data is:  ', len(cv.columns))
print('Count of rows in the data is:  ', len(cv))
```

```
Count of columns in the data is:   2
Count of rows in the data is:   962
```

In [6]:
```python
cv.shape
```

Out[6]: (962, 2)

In [7]:
```python
cv.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 962 entries, 0 to 961
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Category  962 non-null    object
 1   Resume    962 non-null    object
dtypes: object(2)
memory usage: 15.2+ KB
```

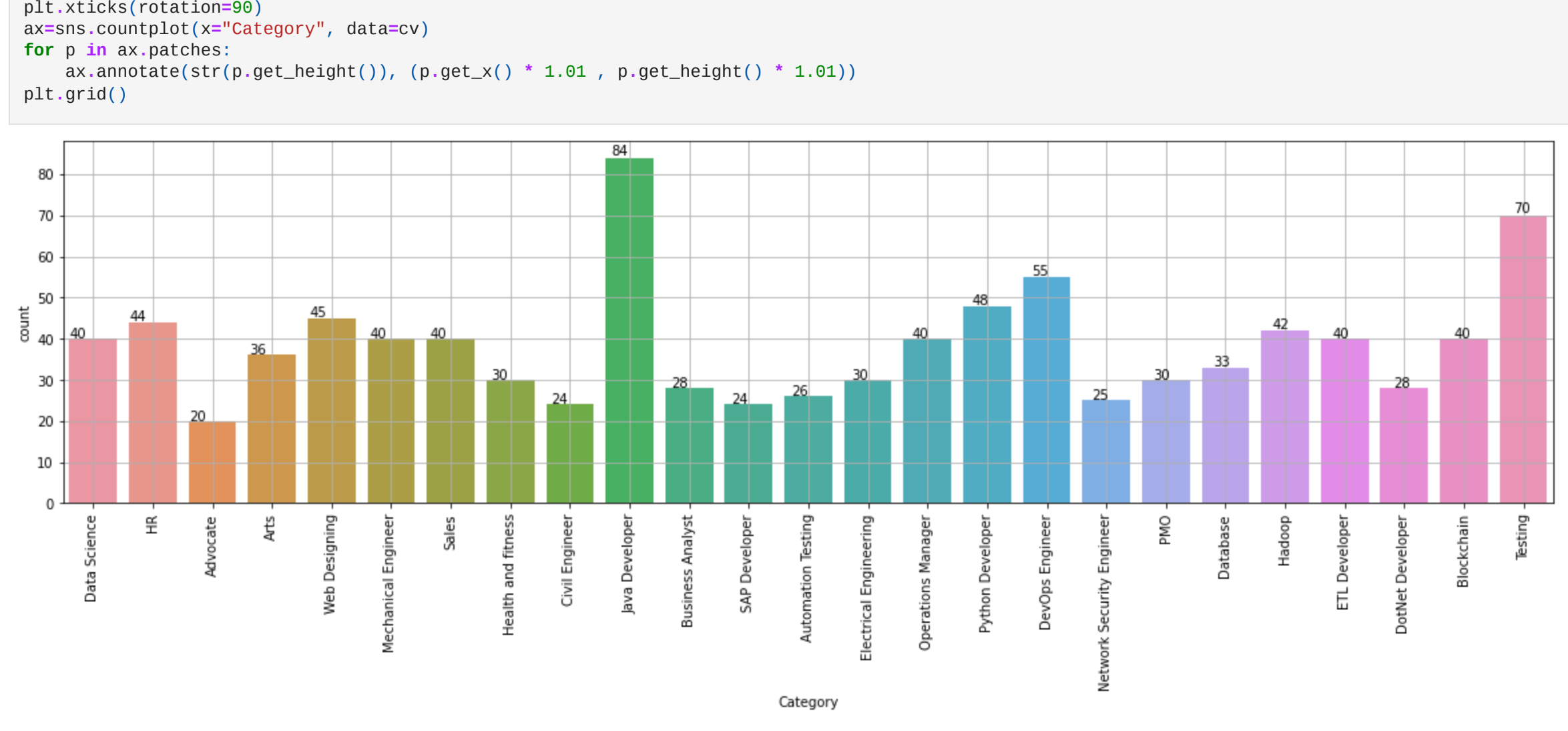
In [8]:
```python
print ("Displaying the distinct categories of resume:\n\n ")
print (cv['Category'].unique())
```

```
Displaying the distinct categories of resume:


['Data Science' 'HR' 'Advocate' 'Arts' 'Web Designing'
 'Mechanical Engineer' 'Sales' 'Health and fitness' 'Civil Engineer'
 'Java Developer' 'Business Analyst' 'SAP Developer' 'Automation Testing'
 'Electrical Engineering' 'Operations Manager' 'Python Developer'
 'DevOps Engineer' 'Network Security Engineer' 'PMO' 'Database' 'Hadoop'
 'ETL Developer' 'DotNet Developer' 'Blockchain' 'Testing']
```

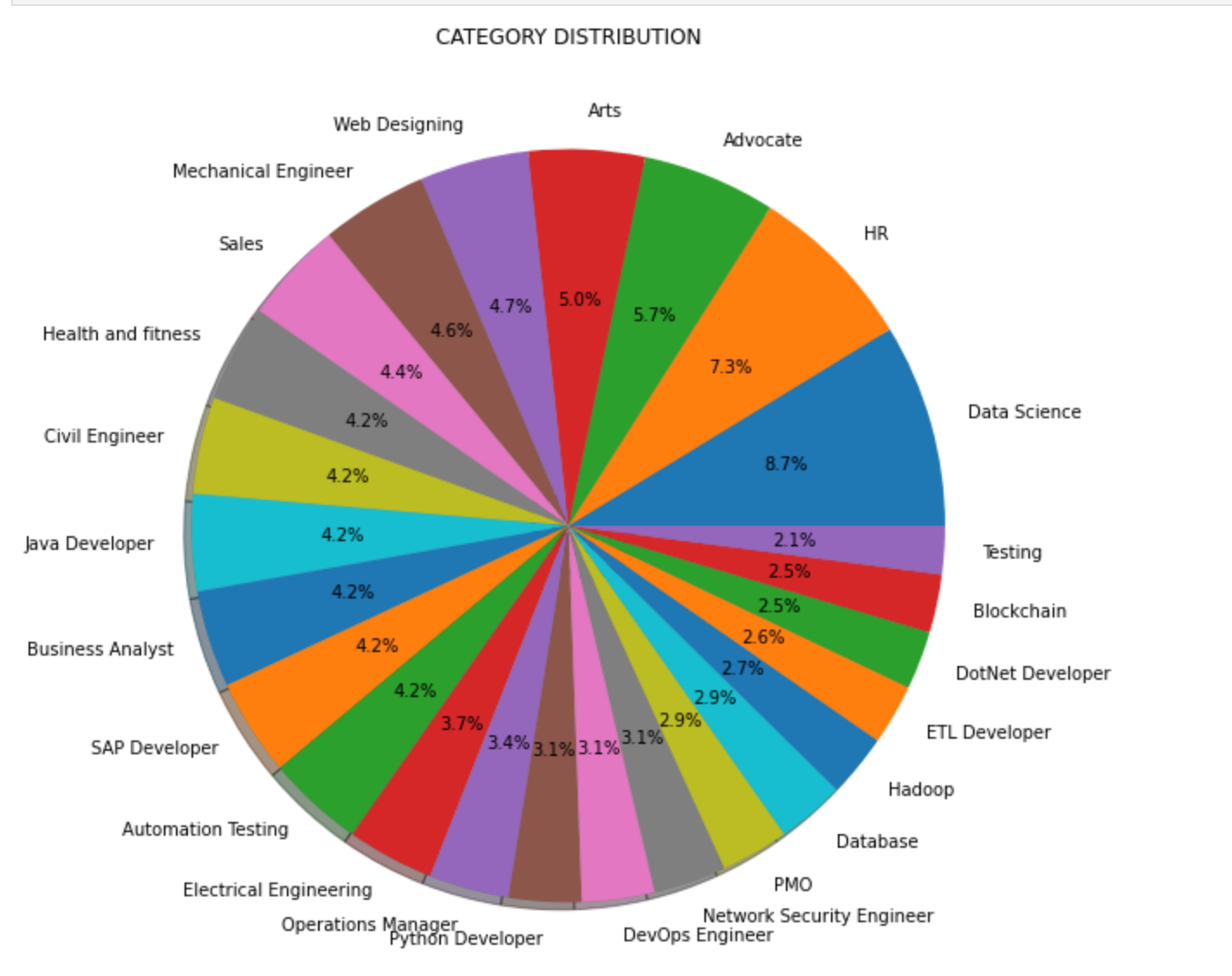
In [9]:
```python
print ("Displaying the distinct categories of resume and the number of records belonging to each category:\n\n")
print (cv['Category'].value_counts())
```

```
Displaying the distinct categories of resume and the number of records belonging to each category:


Java Developer               84
Testing                      70
DevOps Engineer              55
Python Developer             48
Web Designing                45
HR                           44
Hadoop                       42
Operations Manager           40
Sales                        40
Mechanical Engineer          40
Data Science                 40
ETL Developer                40
Blockchain                   40
Arts                         36
Database                     33
PMO                          30
Electrical Engineering       30
Health and fitness           30
DotNet Developer             28
Business Analyst             28
Automation Testing           26
Network Security Engineer    25
Civil Engineer               24
SAP Developer                24
Advocate                     20
Name: Category, dtype: int64
```

In [10]:
```python
plt.figure(figsize=(20,5))
plt.xticks(rotation=90)
ax=sns.countplot(x="Category", data=cv)
for p in ax.patches:
    ax.annotate(str(p.get_height()), (p.get_x() * 1.01 , p.get_height() * 1.01))
plt.grid()
```



In [11]:
```python
targetCounts = cv['Category'].value_counts()
targetLabels  = cv['Category'].unique()

# Make square figures and axes
plt.figure(1, figsize=(22,22))
the_grid = GridSpec(2, 2)


cmap = plt.get_cmap('coolwarm')
plt.subplot(the_grid[0, 1], aspect=1, title='CATEGORY DISTRIBUTION')

source_pie = plt.pie(targetCounts, labels=targetLabels, autopct='%1.1f%%', shadow=True)
plt.show()
```



In [12]:
```python
def cleanResume(resumeText):
    resumeText = re.sub('http\S+\s*', ' ', resumeText)  # remove URLs
    resumeText = re.sub('RT|cc', ' ', resumeText)  # remove RT and cc
    resumeText = re.sub('#\S+', '', resumeText)  # remove hashtags
    resumeText = re.sub('@\S+', '  ', resumeText)  # remove mentions
    resumeText = re.sub('[%s]' % re.escape("""!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~"""), ' ', resumeText)  # remove punctuations
    resumeText = re.sub(r'[^\x00-\x7f]',r' ', resumeText)
    resumeText = re.sub('\s+', ' ', resumeText)  # remove extra whitespace
    return resumeText

cv['cleaned_resume'] = cv.Resume.apply(lambda x: cleanResume(x))
```

In [13]:
```python
cv.head(10)
```

Out[13]:

	Category	Resume	cleaned_resume
0	Data Science	Skills * Programming Languages: Python (pandas...	Skills Programming Languages Python pandas num...
1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E...	Education Details May 2013 to May 2017 B E UIT...
2	Data Science	Areas of Interest Deep Learning, Control Syste...	Areas of Interest Deep Learning Control System...
3	Data Science	Skills â¢ R â¢ Python â¢ SAP HANA â¢ Table...	Skills R Python SAP HANA Tableau SAP HANA SQL ...
4	Data Science	Education Details \r\n MCA YMCAUST, Faridab...	Education Details MCA YMCAUST Faridabad Haryan...
5	Data Science	SKILLS C Basics, IOT, Python, MATLAB, Data Sci...	SKILLS C Basics IOT Python MATLAB Data Science...
6	Data Science	Skills â¢ Python â¢ Tableau â¢ Data Visuali...	Skills Python Tableau Data Visualization R Stu...
7	Data Science	Education Details \r\n B.Tech Rayat and Bahr...	Education Details B Tech Rayat and Bahra Insti...
8	Data Science	Personal Skills â ¢ Ability to quickly grasp t...	Personal Skills Ability to quickly grasp techn...
9	Data Science	Expertise â Data and Quantitative Analysis â...	Expertise Data and Quantitative Analysis Decis...

In [14]:
```python
cv_d=cv.copy()
```

In [15]:
```python
import nltk
import string
from wordcloud import WordCloud
from nltk.corpus import stopwords
```

In [16]:
```python
oneSetOfStopWords = set(stopwords.words('english')+['``',"''"])
totalWords =[]
Sentences = cv['Resume'].values
cleanedSentences = ""
for records in Sentences:
    cleanedText = cleanResume(records)
    cleanedSentences += cleanedText
    requiredWords = nltk.word_tokenize(cleanedText)
    for word in requiredWords:
        if word not in oneSetOfStopWords and word not in string.punctuation:
            totalWords.append(word)

wordfreqdist = nltk.FreqDist(totalWords)
mostcommon = wordfreqdist.most_common(50)
print(mostcommon)
```

```
[('Exprience', 3829), ('months', 3233), ('company', 3130), ('Details', 2967), ('description', 2634), ('1', 2134), ('Project', 1808), ('project', 1579), ('6', 1499), ('data', 1438),
('team', 1424), ('Maharashtra', 1385), ('year', 1244), ('Less', 1137), ('January', 1086), ('using', 1041), ('Skill', 1018), ('Pune', 1016), ('Management', 1010), ('SQL', 990), ('Lt
d', 934), ('management', 927), ('C', 896), ('Engineering', 855), ('Education', 833), ('Developer', 806), ('Java', 773), ('2', 754), ('development', 752), ('monthsCompany', 746),
('Pvt', 730), ('application', 727), ('System', 715), ('reports', 697), ('business', 696), ('India', 693), ('requirements', 693), ('I', 690), ('various', 688), ('A', 688), ('Data',
674), ('The', 672), ('University', 656), ('process', 648), ('Testing', 646), ('test', 638), ('Responsibilities', 637), ('system', 636), ('testing', 634), ('Software', 632)]
```

In [17]:
```python
wc = WordCloud().generate(cleanedSentences)
plt.figure(figsize=(10,10))
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

In [18]:
```python
from sklearn.preprocessing import LabelEncoder

var_mod = ['Category']
le = LabelEncoder()
for i in var_mod:
    cv[i] = le.fit_transform(cv[i])
```

In [19]:
```python
cv.head(10)
```

Out[19]:

	Category	Resume	cleaned_resume
0	6	Skills * Programming Languages: Python (pandas...	Skills Programming Languages Python pandas num...
1	6	Education Details \r\nMay 2013 to May 2017 B.E...	Education Details May 2013 to May 2017 B E UIT...
2	6	Areas of Interest Deep Learning, Control Syste...	Areas of Interest Deep Learning Control System...
3	6	Skills â¢ R â¢ Python â¢ SAP HANA â¢ Table...	Skills R Python SAP HANA Tableau SAP HANA SQL ...
4	6	Education Details \r\n MCA YMCAUST, Faridab...	Education Details MCA YMCAUST Faridabad Haryan...
5	6	SKILLS C Basics, IOT, Python, MATLAB, Data Sci...	SKILLS C Basics IOT Python MATLAB Data Science...
6	6	Skills â¢ Python â¢ Tableau â¢ Data Visuali...	Skills Python Tableau Data Visualization R Stu...
7	6	Education Details \r\n B.Tech Rayat and Bahr...	Education Details B Tech Rayat and Bahra Insti...
8	6	Personal Skills â ¢ Ability to quickly grasp t...	Personal Skills Ability to quickly grasp techn...
9	6	Expertise â Data and Quantitative Analysis â...	Expertise Data and Quantitative Analysis Decis...

In [20]:
```python
cv.Category.value_counts()
```

Out[20]:
```
15    84
23    70
8     55
20    48
24    45
12    44
13    42
6     40
18    40
22    40
3     40
10    40
16    40
1     36
7     33
11    30
14    30
19    30
4     28
9     28
2     26
17    25
21    24
5     24
0     20
Name: Category, dtype: int64
```

In [21]:
```python
cv_d.Category.value_counts() #understanding decode LabelEncoder
```

Out[21]:
```
Java Developer               84
Testing                      70
DevOps Engineer              55
Python Developer             48
Web Designing                45
HR                           44
Hadoop                       42
Operations Manager           40
Sales                        40
Mechanical Engineer          40
Data Science                 40
ETL Developer                40
Blockchain                   40
Arts                         36
Database                     33
PMO                          30
Electrical Engineering       30
Health and fitness           30
DotNet Developer             28
Business Analyst             28
Automation Testing           26
Network Security Engineer    25
Civil Engineer               24
SAP Developer                24
Advocate                     20
Name: Category, dtype: int64
```

In [22]:
```python
del cv_d #clearing the space occupied
```

In [23]:
```python
from scipy.sparse import hstack
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [24]:
```python
requiredText = cv['cleaned_resume'].values
requiredTarget = cv['Category'].values

word_vectorizer = TfidfVectorizer(
    sublinear_tf=True,
    stop_words='english')
word_vectorizer.fit(requiredText)
WordFeatures = word_vectorizer.transform(requiredText)

print ("Feature completed .....")

X_train,X_test,y_train,y_test = train_test_split(WordFeatures,requiredTarget,random_state=42, test_size=0.2,
                                                 shuffle=True, stratify=requiredTarget)
print(X_train.shape)
print(X_test.shape)
```

```
Feature completed .....
(769, 7351)
(193, 7351)
```

In [25]:
```python
clf = OneVsRestClassifier(KNeighborsClassifier())
clf.fit(X_train, y_train)
prediction = clf.predict(X_test)
print('Accuracy of KNeighbors Classifier on training set: {:.2f}'.format(clf.score(X_train, y_train)))
print('Accuracy of KNeighbors Classifier on test set:     {:.2f}'.format(clf.score(X_test, y_test)))
```

```
Accuracy of KNeighbors Classifier on training set: 0.99
Accuracy of KNeighbors Classifier on test set:     0.98
```

In [26]:
```python
print("\n Classification report for classifier %s:\n%s\n" % (clf, metrics.classification_report(y_test, prediction)))
```

```
 Classification report for classifier OneVsRestClassifier(estimator=KNeighborsClassifier()):
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         4
           1       1.00      1.00      1.00         7
           2       0.75      0.60      0.67         5
           3       1.00      1.00      1.00         8
           4       1.00      1.00      1.00         6
           5       1.00      1.00      1.00         5
           6       0.89      1.00      0.94         8
           7       1.00      0.86      0.92         7
           8       1.00      0.91      0.95        11
           9       1.00      1.00      1.00         5
          10       1.00      1.00      1.00         8
          11       0.86      1.00      0.92         6
          12       1.00      1.00      1.00         9
          13       1.00      1.00      1.00         8
          14       1.00      1.00      1.00         6
          15       1.00      1.00      1.00        17
          16       1.00      1.00      1.00         8
          17       1.00      1.00      1.00         5
          18       1.00      1.00      1.00         8
          19       1.00      1.00      1.00         6
          20       1.00      1.00      1.00        10
          21       1.00      1.00      1.00         5
          22       1.00      1.00      1.00         8
          23       0.93      1.00      0.97        14
          24       1.00      1.00      1.00         9

    accuracy                           0.98       193
   macro avg       0.98      0.97      0.97       193
weighted avg       0.98      0.98      0.98       193
```