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

```
df = pd.read_csv('UpdatedResumeDataSet.csv')
```

```
df.head()
```

	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...

Next steps: Generate code with df New interactive sheet

```
df.shape
```

```
(962, 2)
```

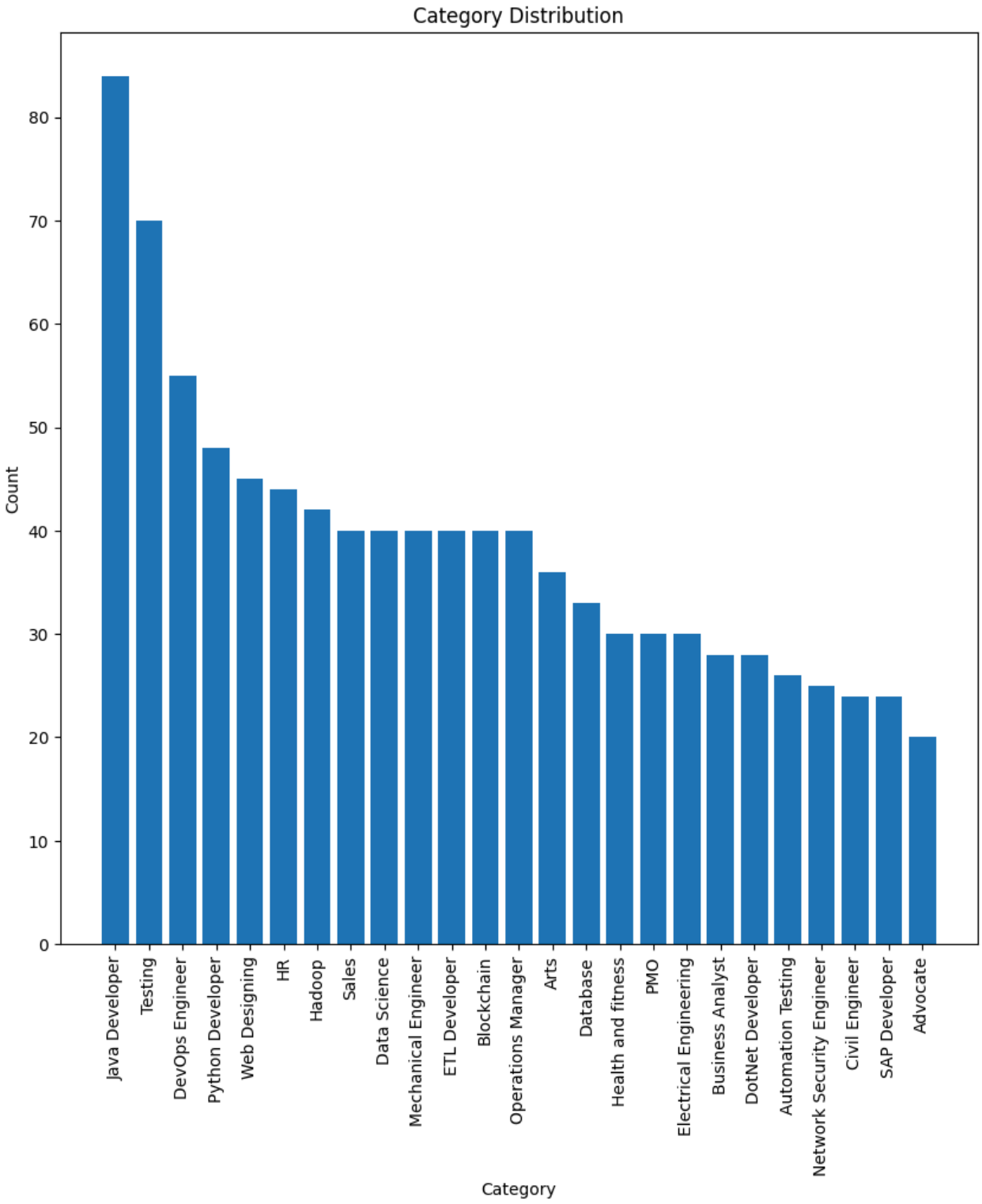
```
df['Category'].value_counts()
```

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

dtype: int64

```
y = df['Category'].value_counts()
x = y.index

plt.figure(figsize=(10,10))
plt.bar(x, y)
plt.xticks(rotation=90)
plt.xlabel("Category")
plt.ylabel("Count")
plt.title("Category Distribution")
plt.show()
```

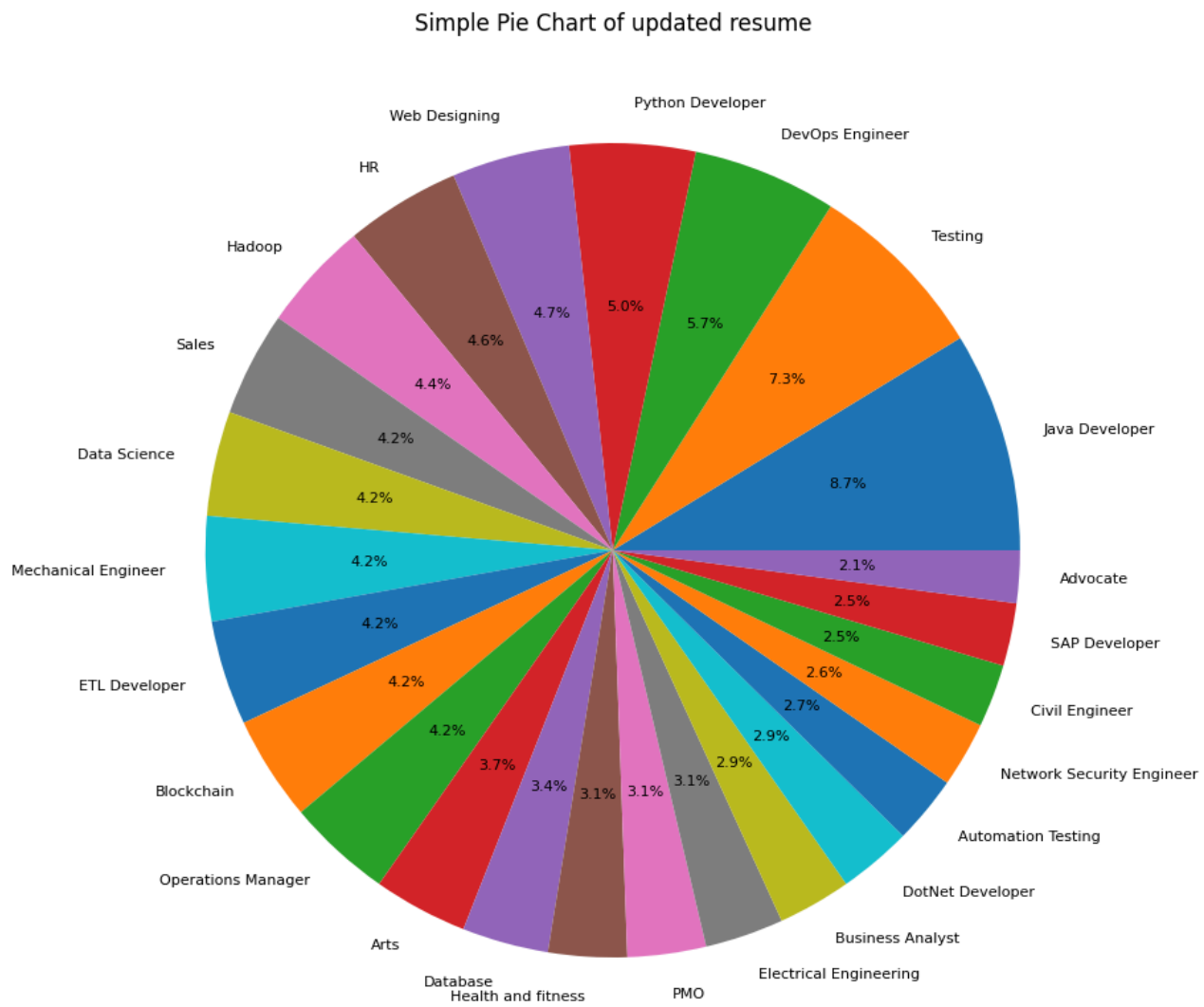



```
df['Category'].unique()
```

```
array(['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'],
      dtype=object)
```

```
plt.figure(figsize=(10,10))
plt.pie(y, labels = x, autopct='%1.1f%%', textprops={'fontsize': 8})
plt.title("Simple Pie Chart of updated resume")
plt.show()
```




```
df['Resume'][0]
```

```
'Skills * Programming Languages: Python (pandas, numpy, scipy, scikit-learn, matplotlib), Sql, Java, Ja
vaScript/JQuery. * Machine learning: Regression, SVM, NaÃ¨ve Bayes, KNN, Random Forest, Decision Trees,
Boosting techniques, Cluster Analysis, Word Embedding, Sentiment Analysis, Natural Language processing,
Dimensionality reduction, Topic Modelling (LDA, NMF), PCA & Neural Nets. * Database Visualizations: Mys
ql, SqlServer, Cassandra, Hbase, ElasticSearch D3.js, DC.js, Plotly, kibana, matplotlib, ggplot, Tablea
u. * Others: Regular Expression, HTML, CSS, Angular 6, Logstash, Kafka, Python Flask, Git, Docker, comp
uter vision - Open CV and understanding of Deep learning.Education Details \r\n\r\nData Science Assuran
ce Associate \r\n\r\nData Science Assurance Associate - Ernst & Young LLP\r\nSkill Details \r\nJAVASCRI
PT- Exprience - 24 months\r\njQuery- Exprience - 24 months\r\nPython- Exprience - 24 monthsCompany Deta
ils \r\ncompany - Ernst & Young LLP\r\ndescription - Fraud Investigation…'
```

```
df['Resume'] = df['Resume'].str.lower()
```

```
import re
import nltk
from nltk.corpus import stopwords

nltk.download('stopwords')
stop_words = set(stopwords.words('english'))

def cleanResume(text):
    cleantext = re.sub(r'http\S+|www\S+', '', text)
    cleantext = re.sub(r'[@#]\w+', '', cleantext)
    cleantext = re.sub(r'[^A-Za-z0-9\s]', '', cleantext)
    cleantext = re.sub(r'[^\w\s]', '', cleantext)

    new_text = []
    for w in cleantext.split():
        if w.lower() in stop_words:
            continue
        else:
            new_text.append(w)

    cleantext = ' '.join(new_text)
    return cleantext
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
df['Resume'] = df['Resume'].apply(cleanResume)
```

```
df['Resume'][0]
```

```
'skills programming languages python pandas numpy scipy scikitlearn matplotlib sql java javascriptjquer
y machine learning regression svm nave bayes knn random forest decision trees boosting techniques clust
er analysis word embedding sentiment analysis natural language processing dimensionality reduction topi
c modelling lda nmf pca neural nets database visualizations mysql sqlserver cassandra hbase elasticsear
ch d3js dcjs plotly kibana matplotlib ggplot tableau others regular expression html css angular 6 logst
ash kafka python flask git docker computer vision open cv understanding deep learningeducation details
data science assurance associate data science assurance associate ernst young llp skill details javascr
ipt exprience 24 months jquery exprience 24 months python exprience 24 monthscompany details company er
nst young llp description fraud investigations dispute services assurance technology assisted review ta
r technology assisted review assists accelerating review process run analy…'
```

```
df.head()
```

	Category	Resume
0	Data Science	skills programming languages python pandas num...
1	Data Science	education details may 2013 may 2017 uitrgpv da...
2	Data Science	areas interest deep learning control system de...
3	Data Science	skills r python sap hana tableau sap hana sql ...
4	Data Science	education details mca ymcaust faridabad haryan...

Next steps: Generate code with df | New interactive sheet

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
le.fit(df['Category'])
df['Category'] = le.transform(df['Category'])
```

```
df.head()
```

	Category	Resume
0	6	skills programming languages python pandas num...
1	6	education details may 2013 may 2017 uitrgpv da...
2	6	areas interest deep learning control system de...
3	6	skills r python sap hana tableau sap hana sql ...
4	6	education details mca ymcaust faridabad haryan...

Next steps: Generate code with df | New interactive sheet

```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer(stop_words='english')
tfidf.fit_transform(df['Resume'])
x = tfidf.transform(df['Resume'])
y = df['Category']
```

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.multiclass import OneVsRestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

knn = OneVsRestClassifier(KNeighborsClassifier())
knn.fit(x_train, y_train)
pred = knn.predict(x_test)

accuracy = accuracy_score(y_test, pred)
precision = precision_score(y_test, pred, average='weighted')
recall = recall_score(y_test, pred, average='weighted')
f1 = f1_score(y_test, pred, average='weighted')

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1 Score:", f1)
```

```
Accuracy: 0.9844559585492227
Precision: 0.9874064478986759
Recall: 0.9844559585492227
F1 Score: 0.9838850508539628
```

```
import pickle
pickle.dump(tfidf, open('tfidf.pkl', 'wb'))
pickle.dump(knn, open('knn.pkl', 'wb'))
```

```
my_resume = """
Abdul Rafay
Email: abdul.rafay@example.com | Phone: +92 300 1234567 | LinkedIn: linkedin.com/in/abdulrafay
Location: Karachi, Pakistan

Professional Summary
Data Scientist with 2 years of experience in analyzing complex datasets, building machine learning models

Skills
```

```
Programming: Python, R, SQL
Machine Learning: Regression, Classification, Clustering, NLP
Libraries: Pandas, NumPy, Scikit-learn, TensorFlow, Keras
Data Visualization: Matplotlib, Seaborn, Tableau, Power BI
Others: Git, Excel, Statistical Analysis

Work Experience
Data Scientist | XYZ Analytics | Jan 2023 – Present
- Developed predictive models to forecast customer churn, improving retention by 15%.
- Built NLP pipelines for sentiment analysis of customer feedback.
- Automated reporting dashboards using Python and Tableau.

Data Analyst | ABC Solutions | Jun 2021 – Dec 2022
- Cleaned and processed large datasets from multiple sources.
- Conducted exploratory data analysis (EDA) to identify trends and patterns.
- Collaborated with cross-functional teams to provide data-driven recommendations.

Education
BS Computer Science | University of Karachi | 2017 – 2021
- Relevant Coursework: Data Mining, Machine Learning, Database Systems

Projects
- Customer Segmentation: Used K-Means clustering to segment customers and optimize marketing strategies.
- Sales Forecasting: Built time-series models to predict monthly sales, reducing stockouts by 10%.
- Resume Screening System: Implemented a TF-IDF + ML model to automatically rank resumes for job positior

Certifications
- Machine Learning by Stanford University (Coursera)
- Data Science Professional Certificate (IBM)
"""
```