Exploratory Data Analysis and visualization of Freelance Platform Dataset

Introduction

This project was assigned by my guide as part of Data analysis with Python, involves an in-depth analysis of a freelance platform dataset available on Kaggle. This dataset pulls the projects posted by clients on PeoplePerHour and made available on Kaggle. The analysis aims to delve into this dataset to uncover valuable insights into the freelance marketplace, providing a practical application of data analytics techniques learned during Data science and analytics essentials.

Dataset Overview

The freelance platform dataset contain information related to freelance projects or job postings along with details about clients and freelancers. Here's a description of the columns present in the dataset:

Category Name: Represents the category or field to which the project belongs (e.g., Web Development, Graphic Design, Writing).

Experience: Indicates the level of experience required or preferred by the client for the freelance job (e.g., Entry-level, Intermediate, Expert).

Sub Category Name: Specifies the sub-category or specific niche within the broader category of the project (e.g., Front-end Development, Logo Design, Content Writing).

Currency: Denotes the currency in which the budget or payment for the project is specified.

Budget: Represents the allocated or proposed budget for the freelance project.

Location: Indicates the location or geographic region associated with the project.

Freelancer Preferred From: Specifies the preferred location or region of the freelancer sought by the client.

Type: Represents the type of project (e.g., Hourly, Fixed-price).

Date Posted: Signifies the date when the project was posted or listed on the freelance platform.

Description: Contains a description or details about the project requirements, scope, or deliverables.

Duration: Indicates the expected or proposed duration or timeline for project completion.

Client Registration Date: Specifies the date when the client registered on the platform.

Client City: Represents the city of the client associated with the project.

Client Country: Denotes the country of the client associated with the project.

Client Currency: Specifies the currency used by the client.

Client Job Title: Represents the job title or designation of the client.

Objectives

- Apply learned data analytics and visualization techniques to explore, clean, and analyze the freelance platform dataset.
- Uncover patterns, trends, and correlations within the dataset to gain insights into client preferences, project characteristics, and market dynamics.
- Create visual representations of the data to effectively communicate findings.
- Utilize the project as a practical showcase of acquired skills and knowledge from the Data science and analytics essentials course curriculum.

Perform necessary EDA on the data

Exploratory Data Analysis (EDA) is a crucial step in data science that involves analyzing and summarizing datasets to uncover patterns, trends, and insights. In EDA we do following steps:

- Observe the dataset
- Find the shape of dataset
- Find any missing values
- Find datatypes

Importing the required libraries for EDA

Below are the libraries that are used in order to perform EDA

```
# import necessary libraries

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

Loading the data into the data frame.

```
# read the dataset

df = pd.read_csv('/content/Freelance Platform Projects.csv') #Read the
CSV file
df.head() # Display the top five rows of the data
```

```
                                    Title         Category Name
Experience  \
0  Banner images for web desgin websites                Design
Entry ($)
1    Make my picture a solid silhouette   Video, Photo & Image
Entry ($)
2                     Bookkeeper needed              Business
Entry ($)
3                    Accountant needed              Business
Entry ($)
4          Guest Post on High DA Website     Digital Marketing  Expert
($$$)
```

```
           Sub Category Name Currency   Budget Location  \
0             Graphic Design      EUR     60.0   remote
1              Image Editing      GBP     20.0   remote
2       Finance & Accounting      GBP     12.0   remote
3  Tax Consulting & Advising      GBP     14.0   remote
4                        SEO      USD  10000.0   remote

  Freelancer Preferred From         Type          Date Posted  \
0                      ALL  fixed_price  2023-04-29 18:06:39
1                      ALL  fixed_price  2023-04-29 17:40:28
2                      ALL  fixed_price  2023-04-29 17:40:06
3                      ALL  fixed_price  2023-04-29 17:32:01
4                      ALL  fixed_price  2023-04-29 17:09:36

                                        Description Duration  \
0  We are looking to improve the banner images on...      NaN
1  Hello \n\nI need a quick designer to make 4 pi...      NaN
2  Hi - I need a bookkeeper to assist with bookke...      NaN
3  Hi - I need an accountant to assist me with un...      NaN
4  Hi, I am currently running a project where I w...      NaN

  Client Registration Date Client City  Client Country Client Currency
\
0               2010-11-03      Dublin          Ireland              EUR

1               2017-02-21      London  United Kingdom              GBP

2               2023-04-09      London  United Kingdom              GBP

3               2023-04-09      London  United Kingdom              GBP

4               2016-07-01      Mumbai           India              USD


    Client Job Title
0     PPC Management
1     Office manager
2          Paralegal
3          Paralegal
4  Guest posts buyer
```

```
# finding size of dataset
df.shape
```

```
(12222, 17)
```

```
# print information about the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12222 entries, 0 to 12221
Data columns (total 17 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   Title                       12222 non-null  object 
 1   Category Name               12222 non-null  object 
 2   Experience                  12222 non-null  object 
 3   Sub Category Name           12222 non-null  object 
 4   Currency                    12222 non-null  object 
 5   Budget                      12222 non-null  float64
 6   Location                    12222 non-null  object 
 7   Freelancer Preferred From   12222 non-null  object 
 8   Type                        12222 non-null  object 
 9   Date Posted                 12222 non-null  object 
 10  Description                 12222 non-null  object 
 11  Duration                    1602 non-null   object 
 12  Client Registration Date    12222 non-null  object 
 13  Client City                 12222 non-null  object 
 14  Client Country              12222 non-null  object 
 15  Client Currency             12222 non-null  object 
 16  Client Job Title            4588 non-null   object 
dtypes: float64(1), object(16)
memory usage: 1.6+ MB
```

Checking the missing or null values.

```
#Return the number of missing values in each column
df.isna().sum()
```

```
Title                            0
Category Name                    0
Experience                       0
Sub Category Name                0
Currency                         0
Budget                           0
Location                         0
Freelancer Preferred From        0
Type                             0
Date Posted                      0
Description                      0
Duration                     10620
Client Registration Date         0
Client City                      0
Client Country                   0
Client Currency                  0
Client Job Title              7634
dtype: int64
```

here are null values in 2 columns i.e. duration and client job title. We will fill these values as not mentioned.

```
# Fill null values
df['Duration'].replace(np.nan,'Not mentioned',inplace=True) #fill null
values

df['Client Job Title'].replace(np.nan,'Not mentioned',inplace=True)
#fill null values

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12222 entries, 0 to 12221
Data columns (total 17 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Title                      12222 non-null  object
 1   Category Name              12222 non-null  object
 2   Experience                 12222 non-null  object
 3   Sub Category Name          12222 non-null  object
 4   Currency                   12222 non-null  object
 5   Budget                     12222 non-null  float64
 6   Location                   12222 non-null  object
 7   Freelancer Preferred From  12222 non-null  object
 8   Type                       12222 non-null  object
 9   Date Posted                12222 non-null  object
 10  Description                12222 non-null  object
 11  Duration                   12222 non-null  object
 12  Client Registration Date   12222 non-null  object
 13  Client City                12222 non-null  object
 14  Client Country             12222 non-null  object
 15  Client Currency            12222 non-null  object
 16  Client Job Title           12222 non-null  object
dtypes: float64(1), object(16)
memory usage: 1.6+ MB
```

```
df.columns
```

```
Index(['Title', 'Category Name', 'Experience', 'Sub Category Name',
'Currency',
       'Budget', 'Location', 'Freelancer Preferred From', 'Type',
       'Date Posted', 'Description', 'Duration', 'Client Registration
Date',
       'Client City', 'Client Country', 'Client Currency', 'Client Job
Title'],
      dtype='object')
```

The Budget is given in 3 types of currencies. so we will convert it into one i.e usd.

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

```
array(['EUR', 'GBP', 'USD'], dtype=object)
```

```
def convert_to_usd(data):
    conversion = {'EUR': 1.07, 'GBP': 1.24, 'USD': 1}
    data['Budget'] = [data.loc[i,'Budget'] *
conversion[data.loc[i,'Currency']] for i in data.index]
    data.drop(columns=['Currency'] , axis=1, inplace=True)
    return data

df = convert_to_usd(df)
df.head()
```

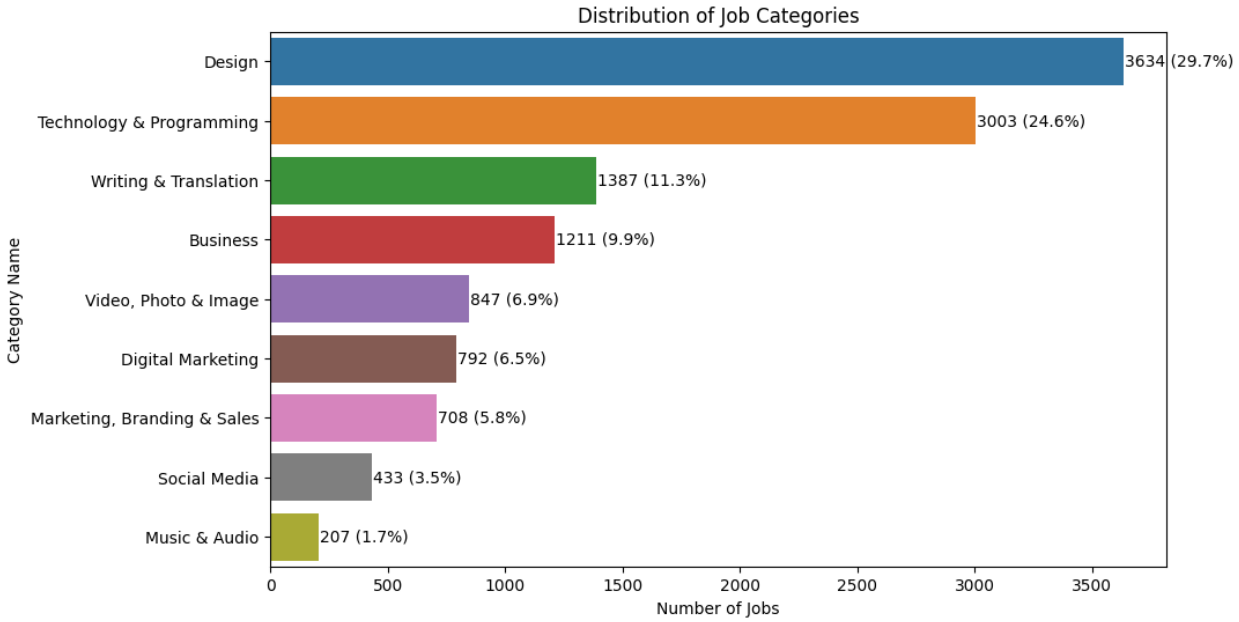
```
                                    Title         Category Name
Experience  \
0  Banner images for web desgin websites                 Design
Entry ($)
1    Make my picture a solid silhouette   Video, Photo & Image
Entry ($)
2                      Bookkeeper needed              Business
Entry ($)
3                     Accountant needed              Business
Entry ($)
4          Guest Post on High DA Website     Digital Marketing  Expert
($$$)

           Sub Category Name    Budget Location Freelancer Preferred
From  \
0             Graphic Design     64.20   remote
ALL
1              Image Editing     24.80   remote
ALL
2       Finance & Accounting     14.88   remote
ALL
3  Tax Consulting & Advising     17.36   remote
ALL
4                        SEO  10000.00   remote
ALL

          Type          Date Posted  \
0  fixed_price  2023-04-29 18:06:39
1  fixed_price  2023-04-29 17:40:28
2  fixed_price  2023-04-29 17:40:06
3  fixed_price  2023-04-29 17:32:01
4  fixed_price  2023-04-29 17:09:36

                                          Description       Duration  \
0  We are looking to improve the banner images on...  Not mentioned
1  Hello \n\nI need a quick designer to make 4 pi...  Not mentioned
```

```
2  Hi - I need a bookkeeper to assist with bookke...  Not mentioned
3  Hi - I need an accountant to assist me with un...  Not mentioned
4  Hi, I am currently running a project where I w...  Not mentioned

  Client Registration Date Client City  Client Country Client Currency
\
0               2010-11-03      Dublin         Ireland              EUR

1               2017-02-21      London  United Kingdom              GBP

2               2023-04-09      London  United Kingdom              GBP

3               2023-04-09      London  United Kingdom              GBP

4               2016-07-01      Mumbai           India              USD


    Client Job Title
0     PPC Management
1     Office manager
2          Paralegal
3          Paralegal
4  Guest posts buyer
```

Visualizations

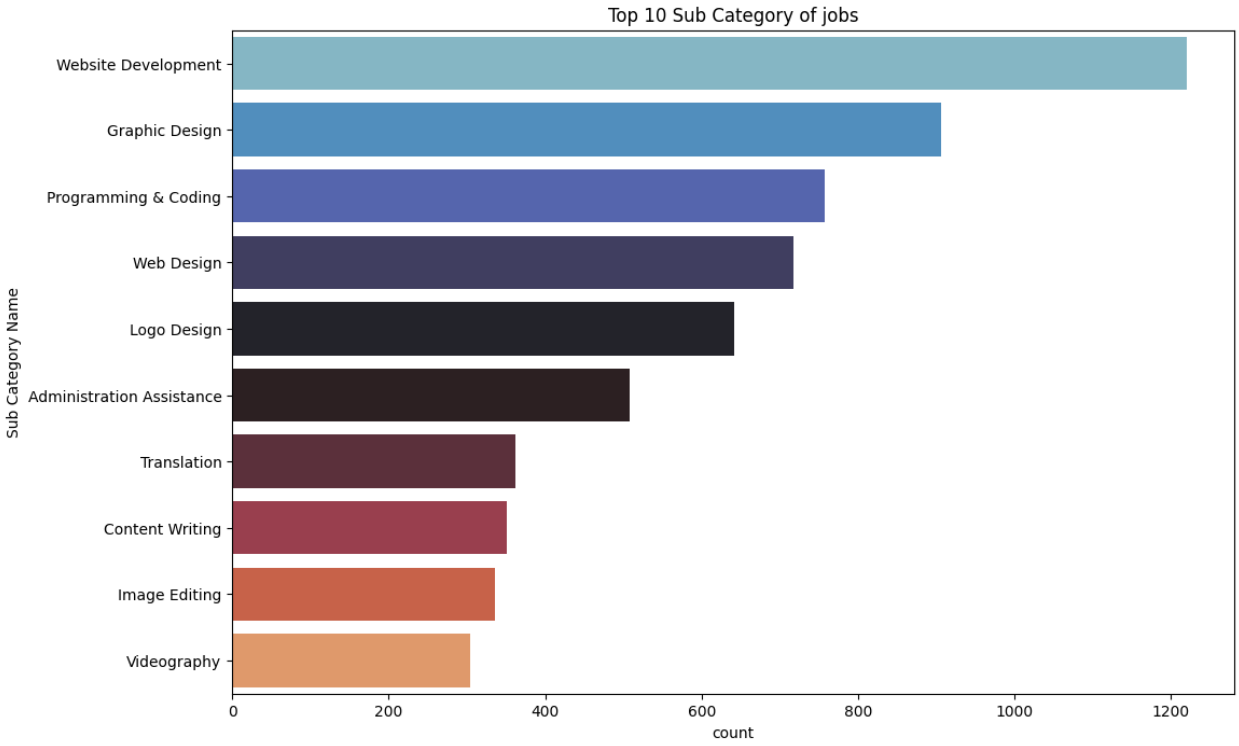
1.Categories of jobs available on freelance platform.

```
# Calculate value counts and percentages for 'Category Name'
category_counts = df['Category Name'].value_counts()
category_percentages = df['Category
Name'].value_counts(normalize=True) * 100

plt.figure(figsize=(10, 6))
sns.countplot(y='Category Name', data=df, order=category_counts.index)
plt.title('Distribution of Job Categories')
plt.xlabel('Number of Jobs')
plt.ylabel('Category Name')

# Show counts and percentages on the plot
for i, count in enumerate(category_counts):
    plt.text(count + 5, i, f'{count} ({category_percentages[i]:.1f}
%)', va='center')

plt.show()
```



2. The top-10 jobs sub-categories on freelance platfrom.

```
plt.figure(figsize = (12,8))
sns.countplot(y='Sub Category Name',order = df['Sub Category
Name'].value_counts().index[0:10],data = df,palette='icefire')
plt.title('Top 10 Sub Category of jobs')

Text(0.5, 1.0, 'Top 10 Sub Category of jobs')
```



3.Budget cost for each job category.

```
categories = df.groupby(['Category Name'])
['Budget'].sum().reset_index()
fig = px.bar(categories,x='Category Name',y='Budget',title='Budget for
Each Category',
                 color="Category Name",height = 700)
fig.show()
```

4. Distribution of Experience level for freelancers.

```
type_count = list(df['Experience'].value_counts())
colors = ['14213D','FCA311']
type_ls = list(df['Experience'].value_counts().index)
fig = px.pie(values=type_count,names=type_ls)
fig.update_layout(title_text='Distribution of Experience level for
freelancers', title_x=0.5,height = 600)
fig.update_traces(rotation=90,textposition='inside',textinfo='label+pe
rcent+value',marker=dict(colors=colors))
```

5.Distribution of clients across countries.

```
Client_Country = list(df['Client Country'].unique())
Client_Country_count = list(df['Client Country'].value_counts())
fig = px.pie(df,values=Client_Country_count,names=Client_Country,
hole=0.5,color_discrete_sequence=px.colors.sequential.RdBu)
fig.update_layout(title_text='Country wise number of clients ',
title_x=0.5)
fig.update_traces(textposition='inside',textinfo='label+percent+value'
)
fig.show()
```

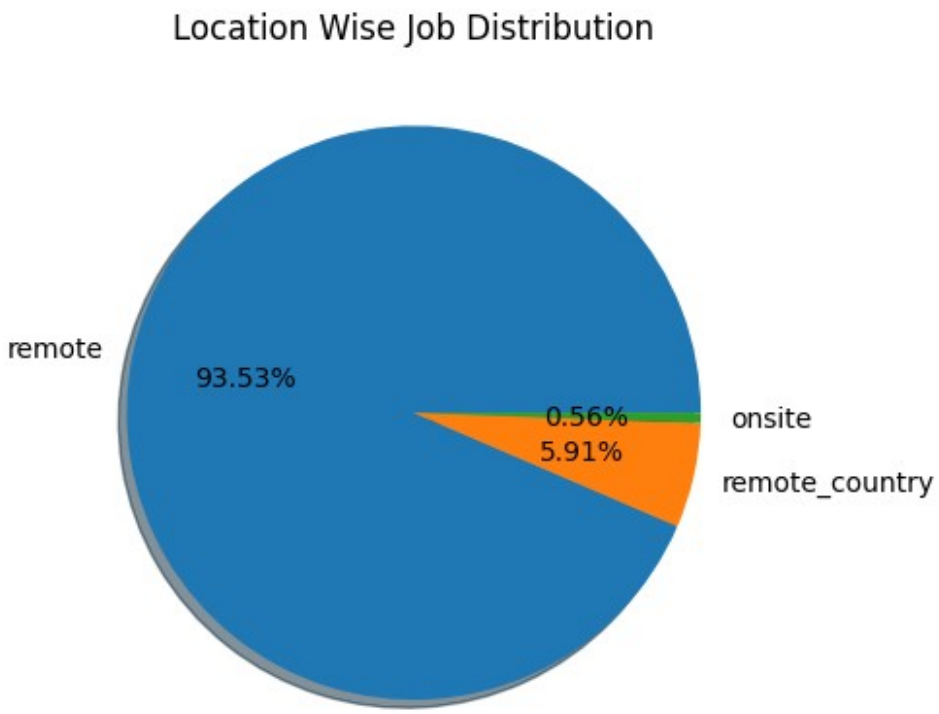
6.How many number of clients registered every year?

```
 #Convert 'Client Registration Date' column to datetime format
df['Client Registration Date'] = pd.to_datetime(df['Client
Registration Date'])

# Extract year from the 'Client Registration Date' column
df['Registration Year'] = df['Client Registration Date'].dt.year

# Count the number of clients registered each year
clients_registered_per_year = df['Registration
Year'].value_counts().sort_index()

print("Number of clients registered each year:")
print(clients_registered_per_year)
```

```
Number of clients registered each year:
2007       2
2008      58
2009      76
2010     277
2011     270
2012     350
2013     532
2014     568
2015     716
2016     736
2017     815
2018     733
2019     794
2020     996
2021     732
2022    1134
2023    3433
Name: Registration Year, dtype: int64
```

7. What are the preffered job Locations?

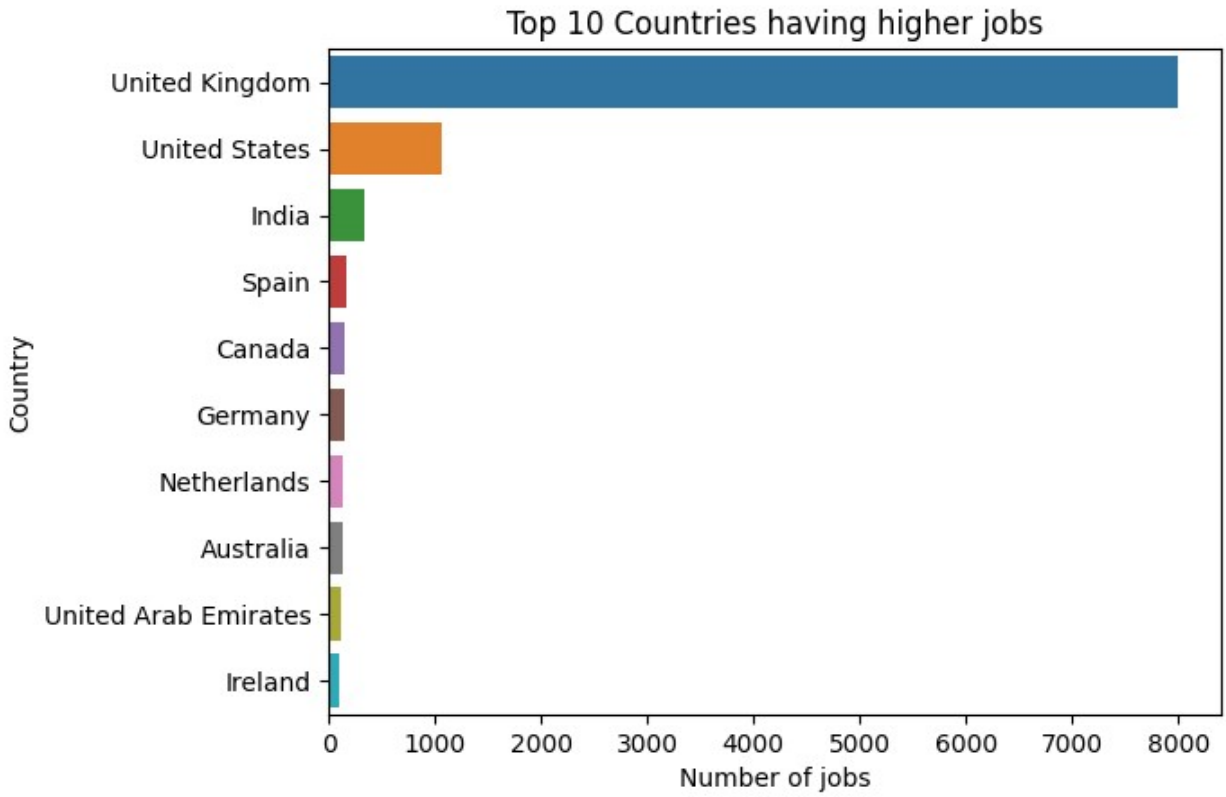
```
location_counts=df["Location"].value_counts()
plt.pie(location_counts,labels=location_counts.index,autopct='%0.2f%
%',shadow=True)
plt.title("Location Wise Project Distribution")
plt.show()
```



8. Top 10 contries having higher jobs.

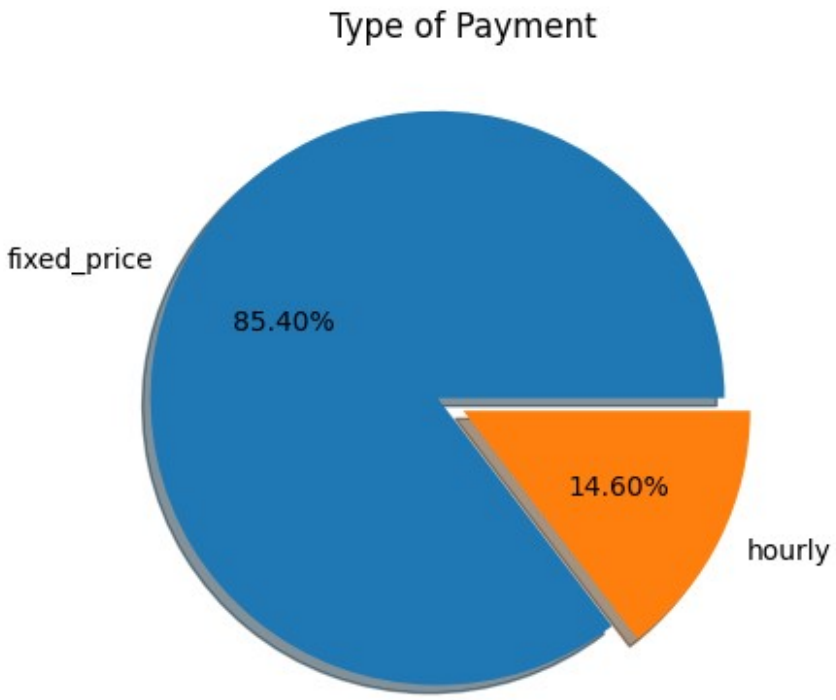
```
top=df["Client Country"].value_counts().sort_values(ascending=False)
top_new=top.head(10)
sns.barplot(x=top_new.values,y=top_new.index)
plt.title("Top 10 Countries having higher jobs")
plt.xlabel("Number of jobs")
plt.ylabel("Country")
```

```
Text(0, 0.5, 'Country')
```



9. Types of payments for freelancers.

```
grp=df["Type"].value_counts()
ex=[0.1,0]
plt.pie(grp,explode=ex,labels=grp.index,autopct='%0.2f%%',shadow=True)
plt.title("Type of Payment")
plt.show()
```



Conclusion

1. The freelance platform dataset has 17 columns and 12222 rows. The Budget is in 3 currencies i.e.'EUR', 'GBP', 'USD', so I converted these in one i.e. USD. for better understanding.
2. Different job categories available on freelance platform are 'Design', 'Business', 'Digital marketing', 'Marketing, Branding and Sales', 'Music $ Audio', 'Social media', 'Technology & programming', 'Video, Photo, Image', 'Writing & translation'. Design category is the most popular, while the Music & Audio category has the fewest jobs.
3. Website developement is a subcategory having highest number of jobs.
4. Technology & Programming, Writing & Translation, Video, Photo & Image, and Social Media are the job categories having higher higher budgets.
5. Expert and Entry levels are in high demand, while there are fewer opportunities for those with Intermediate experience.
6. 65.5% of the total clients are from only Ireland.
7. Number of clients are increasing every year.
8. Remote locations have a higher concentration of freelancing jobs.
9. 93.53% of these job opportunities provide the flexibility for employees to work remotely.
10. United Kingdom, the United States, and India are the countries having highest job opportunities.
11. The majority of jobs (85.40%) offer a fixed payment for the completion of a specific task or project. Conversely, 14.60% of jobs offer an hourly payment structure, where employees are compensated based on the number of hours worked.