```
import numpy as np
import pandas as pd
import seaborn as sns
```

```
df = pd.read_csv("/content/startup_funding.csv")
```

```
df.isnull().sum()
```

```
Sr No                   0
Date dd/mm/yyyy         0
Startup Name            0
Industry Vertical     171
SubVertical           936
City  Location        180
Investors Name         24
InvestmentnType         4
Amount in USD         960
Remarks              2625
dtype: int64
```

Handle missing values and Data Preprocessing

```
df.drop(["Sr No", "SubVertical", "Remarks"],axis=1, inplace=True)
```

```
df['City  Location'] = df['City  Location'].str.replace(r'Bangalore', 'Bengaluru', regex=True)
df['City  Location'] = df['City  Location'].str.replace(r'Bhubneswar', 'Bhubaneswar', regex=True)
df['City  Location'] = df['City  Location'].str.replace(r'Kolkatta', 'Kolkata', regex=True)
df['City  Location'] = df['City  Location'].str.replace(r'Nw Delhi', 'New Delhi', regex=True)
df['City  Location'] = df['City  Location'].str.replace(r'\bUS\b', 'USA', regex=True)
df['City  Location'] = df['City  Location'].str.replace(r'\\\\xc2\\\\xa0', '', regex=True)
df['City  Location'] = df['City  Location'].fillna('')  # Fill NaN values with an empty string
df['City  Location'] = df['City  Location'].replace({'Ahemadabad': 'Ahmedabad', 'Ahemdabad': 'Ahmedabad'})
df.loc[df['City  Location'].str.contains('/'), 'City  Location'] = 'Multiple Cities'
df.loc[df['City  Location'].str.contains('&'), 'City  Location'] = 'Multiple Cities'
df.loc[df['City  Location'].str.contains('and'), 'City  Location'] = 'Multiple Cities'
df.loc[df['City  Location'].str.contains(','), 'City  Location'] = 'Multiple Cities'
```

```
df['City  Location'].fillna("Unknown", inplace = True)
df['Industry Vertical'].fillna("Unknown", inplace = True)
df['Investors Name'].fillna("Unknown",inplace=True)
df["InvestmentnType"].fillna(pd.Series(np.random.choice(["Seed Funding", "Private Equity"], size=len(df.index))), inp
```

```
df["Amount in USD"] = df["Amount in USD"].str.replace(r"\\\\xc2\\\\xa0", "", regex=True)
df["Amount in USD"] = df["Amount in USD"].str.replace(r"\\xc2\\xa0", "", regex=True)
df["Amount in USD"] = df["Amount in USD"].str.replace(r"unknown", "undisclosed", regex=True)
df["Amount in USD"] = df["Amount in USD"].str.replace(r"Undisclosed", "undisclosed", regex=True)
df["Amount in USD"] = df["Amount in USD"].str.replace(r"N/A", "undisclosed", regex=True)
df["Amount in USD"] = df["Amount in USD"].str.replace(r"\+", "", regex=True)
df["Amount in USD"].fillna("undisclosed", inplace=True)
```

```
# mean = df['Amount in USD'].mean()
# median = df['Amount in USD'].median()
# mode = df['Amount in USD'].mode()
# print(f"Mean:  {mean} , Median: {median} , Mode: {mode}")
```

```
# df['Amount in USD'].fillna(df['Amount in USD'].median(), inplace=True)
```

```
df.isnull().sum()
```

```
Date dd/mm/yyyy      0
Startup Name         0
Industry Vertical    0
City  Location       0
Investors Name       0
InvestmentnType      0
Amount in USD        0
dtype: int64
```

Transform variables using log transformation or standardization.

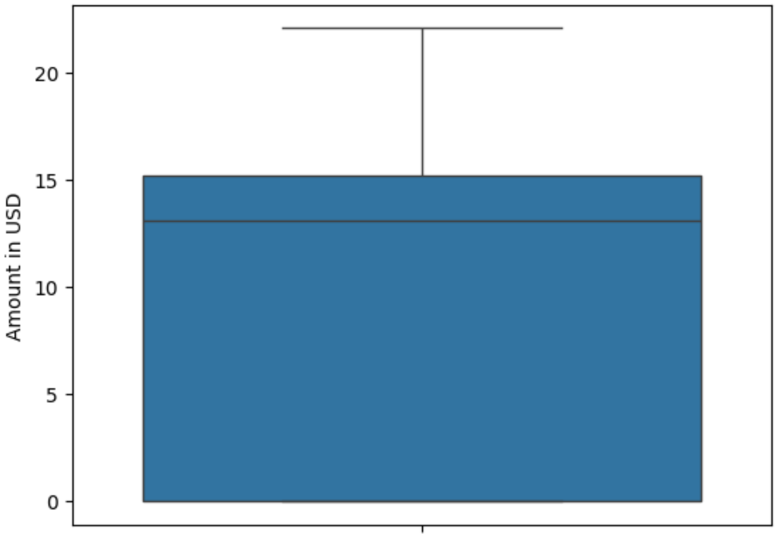
```
df["Amount in USD"] = df["Amount in USD"].str.replace(r",", "", regex=True)
df["Amount in USD"] = df["Amount in USD"].str.replace(r"undisclosed", "1", regex=True)
df["Amount in USD"] = df["Amount in USD"].astype("float64")
df["Amount in USD"] = np.log(df["Amount in USD"])
df.head()
```

	Date dd/mm/yyyy	Startup Name	Industry Vertical	City Location	Investors Name	InvestmentnType	Amount in USD
0	09/01/2020	BYJU'S	E-Tech	Bengaluru	Tiger Global Management	Private Equity Round	19.113828
1	13/01/2020	Shuttl	Transportation	Gurgaon	Susquehanna Growth Equity	Series C	15.900983
2	09/01/2020	Mamaearth	E-commerce	Bengaluru	Sequoia Capital India	Series B	16.725623
3	02/01/2020	https://www.wealthbucket.in/	FinTech	New Delhi	Vinod Khatumal	Pre-series A	14.914123
4	02/01/2020	Fashor	Fashion and Apparel	Mumbai	Sprout Venture Partners	Seed Round	14.403297

Removing outliers based on a specific threshold.

```
threshold = 1.5

# Calculate the IQR
Q1 = df['Amount in USD'].quantile(0.25)
Q3 = df['Amount in USD'].quantile(0.75)
IQR = Q3 - Q1

# Define the lower and upper bounds to identify outliers
lower_bound = Q1 - threshold * IQR
upper_bound = Q3 + threshold * IQR

# Filter the DataFrame to remove outliers
df_filtered = df[(df['Amount in USD'] >= lower_bound) & (df['Amount in USD'] <= upper_bound)]
df_filtered.tail()

sns.boxplot(df_filtered['Amount in USD'])
```

```
<Axes: ylabel='Amount in USD'>
```



Remove duplicate records from a data frame

```
df = df.drop_duplicates()
```

```
df.reset_index()
```

	index	Date dd/mm/yyyy	Startup Name	Industry Vertical	City Location	Investors Name	InvestmentnType	Amount in USD
0	0	09/01/2020	BYJU’S	E-Tech	Bengaluru	Tiger Global Management	Private Equity Round	19.113828
1	1	13/01/2020	Shuttl	Transportation	Gurgaon	Susquehanna Growth Equity	Series C	15.900983
2	2	09/01/2020	Mamaearth	E-commerce	Bengaluru	Sequoia Capital India	Series B	16.725623
3	3	02/01/2020	https://www.wealthbucket.in/	FinTech	New Delhi	Vinod Khatumal	Pre-series A	14.914123
4	4	02/01/2020	Fashor	Fashion and Apparel	Mumbai	Sprout Venture Partners	Seed Round	14.403297
...	...	...	...	...	...	...	...	...
3039	3039	29/01/2015	Printvenue	Unknown		Asia Pacific Internet Group	Private Equity	15.319588
3040	3040	29/01/2015	Graphene	Unknown		KARSEMVEN Fund	Private Equity	13.623139
3041	3041	30/01/2015	Mad Street Den	Unknown		Exfinity Fund, GrowX Ventures.	Private Equity	14.220976
3042	3042	30/01/2015	Simplotel	Unknown		MakeMyTrip	Private Equity	0.000000
						UK based Group of		