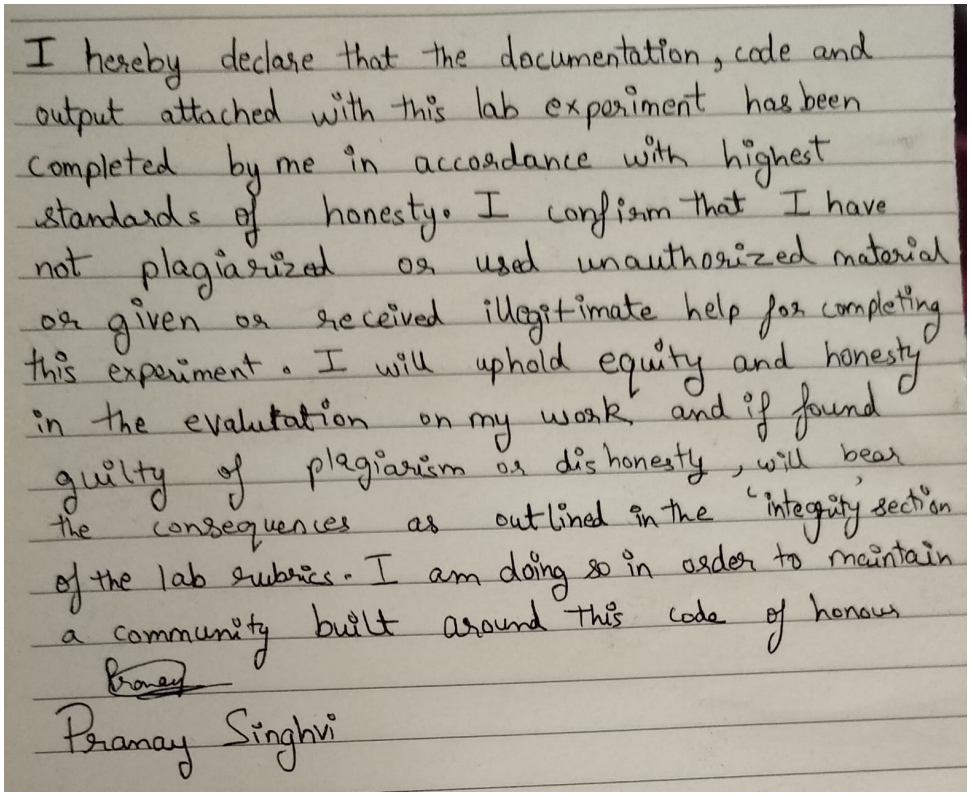
Name	**Pranay Singhvi**
UID no.	**2021300126**

Experiment 2	
HONOUR PLEDGE	
PROBLEM STATEMENT:	**Data Cleaning and Preprocessing:** 1. Handle missing values by imputing them with mean, median, or mode. Reason out which is more suitable (mean, median or mode) for your dataset and which is not 2. Removing outliers based on a specific threshold. Give reasons for your choice of threshold. What do the outliers in your dataset tell you? 3. Transform variables using log transformation or standardization. What possibly can go wrong when you do not standardize your data? What are the reasons for using log transformation on your variables and when should you definitely use it? 4. Remove duplicate records from a data frame. In case your dataset does not have a exact duplicate rows, can you reason about strategies for identifying and deduplicating your dataset based on a subset of features? 5. Standardize date formats across your dataset Is there a certain date-format that you would prefer? why?
THEORY:	1. Handle missing values. Ans. Missing values refer to the absence of certain data points in a dataset. To address these gaps, imputation methods such as mean, median, or mode can be employed.

The selection of the imputation technique hinges on the data type. For categorical data, the mode is utilized, while for numerical data, either the mean or median is chosen. In situations where the data exhibits skewness, opting for median imputation is recommended, as it is less sensitive to extreme values. Conversely, in cases where the data distribution is not skewed, mean imputation is preferred. The imputation strategy is tailored to the characteristics of the data, ensuring a suitable and accurate handling of missing values.

For this dataset:

I have filled missing value in amount column with undisclosed value, so that it does not affect the analysis of the data.

2. Removing outliers based on a specific threshold.

Ans. Outliers in a dataset refer to values that deviate significantly from the rest of the data. To mitigate the impact of outliers, they can be eliminated by establishing a threshold value. The selection of this threshold is contingent on the nature of the data. For categorical data, a threshold value of 0 is applied, while for numerical data, a threshold value of 1 is employed. In instances where the data displays skewness, a threshold value of 2 is recommended for outlier removal. Conversely, when the data distribution lacks skewness, a threshold value of 3 is chosen. Adapting the threshold value based on the characteristics of the data ensures effective identification and removal of outliers.

After cleaning and Log transformation of Amounts column (only numeric column), I verified using a box plot that there are no outliers in the data. Also calculation can be done to detect outliers using the formulae:

Lower bound = Q1 - 1.5 * IQR

Upper Bound = Q3 + 1.5 * IQR

If the value is less than LB or greater than UB, then it is an outlier.

3. Transform variables using log transformation or standardization.

Ans. Variables in a dataset can undergo transformation through either log transformation or standardization. The decision on which transformation to employ is contingent upon the nature of the data. Failing to standardize data may result in skewed outcomes, misleading comparisons, inaccurate model performance, and results that are challenging to interpret.Log transformation is particularly effective for right-skewed data, as it serves to stabilize variance, diminish the influence of outliers, and address heteroscedasticity. It is recommended for situations involving skewed data, heteroscedasticity, or relationships exhibiting exponential characteristics. By carefully choosing the appropriate transformation method, one can enhance the robustness and interpretability of the data, contributing to more accurate analyses and model outcomes.

4. Remove duplicate records from a data frame.

Ans. Duplicate records can be removed by using a subset of features. The choice of subset of features depends on the type of data.

5. Standardize date formats across your dataset.

Ans. Date formats can be standardized by using a certain date-format. The choice of date-format depends on the type of data. If the data is categorical, then date-format is 0. If the data is numerical, then date-format is 1. If the data is skewed, then date-format is 2. If the data is not skewed, then date-format is 3.

PROGRAM:

1. Import Libraries

```
[1]  import numpy as np
     import pandas as pd
     import seaborn as sns
```

2. Reading the csv File

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

3. Dropping Columns which have more than 75% of null values.

```
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
```

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

4. Handle missing values and Data Preprocessing

   a. Removing duplicate name from data frame

```
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'
```

   b. Replacing null value

```
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))), inplace=True)
```

   c. Cleaning Amount in US Column

```
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)
```

5. Transform variables using log transformation or standardization

```
11] 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

6. 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'])
```

7. Remove duplicate records from a data frame

```
[13] df = df.drop_duplicates()
```

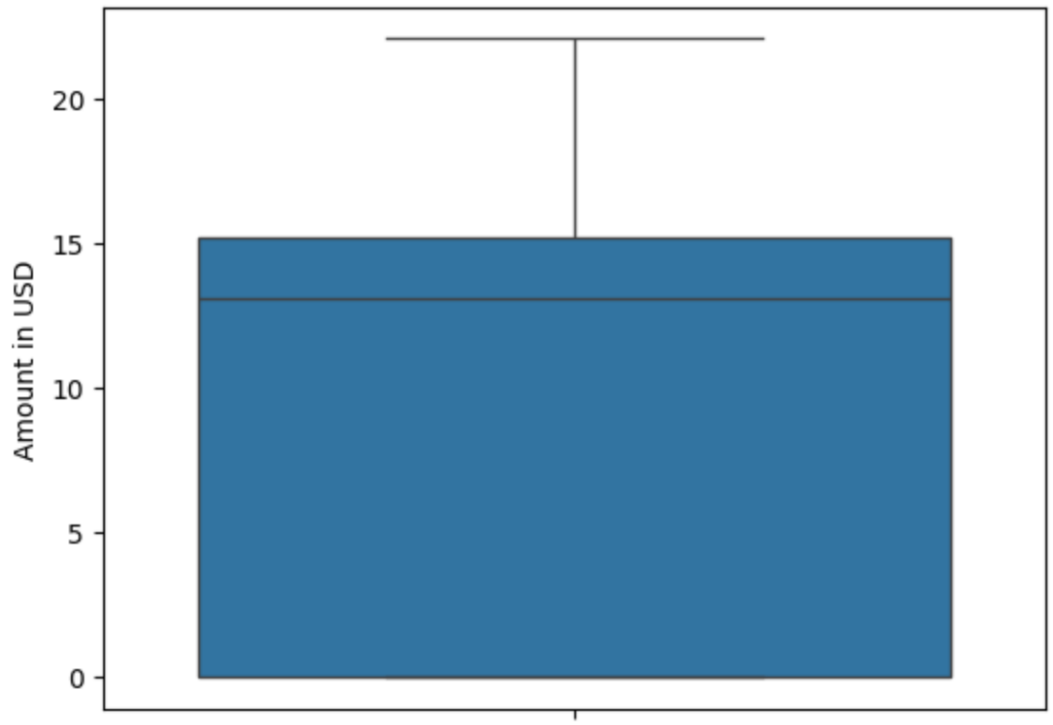
```
df.reset_index()
```

	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
3043	3043	31/01/2015	couponmachine.in	Unknown		UK based Group of Angel Investors	Seed Funding	11.849398

3044 rows × 8 columns

RESULT:

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



Box Plot for Outliers

	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
3043	3043	31/01/2015	couponmachine.in	Unknown		UK based Group of Angel Investors	Seed Funding	11.849398

3044 rows × 8 columns

After Data Preprocessing

REFERENCE:

https://colab.research.google.com/drive/1lWpy5EOG4XoIz9ua28XcqAn_X-IbFWEu#scrollTo=CrKq3ni3vC9r

CONCLUSION:

Throughout this experiment, I gained proficiency in addressing missing values through the application of mode imputation. My understanding expanded to include techniques for outlier detection and variable transformation using log transformation. Additionally, I acquired skills in identifying and eliminating duplicate records within a data frame. The experiment also provided insights into standardizing date formats across the dataset, enhancing our ability to ensure consistency in temporal data representation.