Importing File from local to google_colab

```
from google.colab import files
uploaded = files.upload()
```

Choose Files top_insta_i…ers_data.csv

- **top_insta_influencers_data.csv**(text/csv) - 12628 bytes, last modified: 7/16/2025 - 100% done

Saving top_insta_influencers_data.csv to top_insta_influencers_data.csv

Reading Imported data file

```
import pandas as pd

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

```
df.head()
```

	rank	channel_info	influence_score	posts	followers	avg_likes	60_day_eng_rate	new_post_avg_like	total_likes	country
0	1	cristiano	92	3.3k	475.8m	8.7m	1.39%	6.5m	29.0b	Spain
1	2	kyliejenner	91	6.9k	366.2m	8.3m	1.62%	5.9m	57.4b	United States
2	3	leomessi	90	0.89k	357.3m	6.8m	1.24%	4.4m	6.0b	NaN
3	4	selenagomez	93	1.8k	342.7m	6.2m	0.97%	3.3m	11.5b	United States
4	5	therock	91	6.8k	334.1m	1.9m	0.20%	665.3k	12.5b	United States

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
df.describe()
```

	rank	influence_score
count	200.000000	200.000000
mean	100.500000	81.820000
std	57.879185	8.878159
min	1.000000	22.000000
25%	50.750000	80.000000
50%	100.500000	84.000000
75%	150.250000	86.000000
max	200.000000	93.000000

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   rank               200 non-null    int64
 1   channel_info       200 non-null    object
 2   influence_score    200 non-null    int64
 3   posts              200 non-null    object
 4   followers          200 non-null    object
 5   avg_likes          200 non-null    object
 6   60_day_eng_rate    200 non-null    object
 7   new_post_avg_like  200 non-null    object
 8   total_likes        200 non-null    object
 9   country            138 non-null    object
dtypes: int64(2), object(8)
memory usage: 15.8+ KB
```

Finding null values for better insights

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

	0
rank	0
channel_info	0
influence_score	0
posts	0
followers	0
avg_likes	0
60_day_eng_rate	0
new_post_avg_like	0
total_likes	0
country	62

dtype: int64

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

```
np.int64(62)
```

Finding Duplicate values

```
df.duplicated().sum()
```

```
np.int64(0)
```

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

```
array(['Spain', 'United States', nan, 'Canada', 'Brazil', 'Netherlands',
       'United Kingdom', 'India', 'Uruguay', 'Turkey', 'Indonesia',
       'Colombia', 'France', 'Australia', 'Italy', 'United Arab Emirates',
       'Puerto Rico', "CÃ´te d'Ivoire", 'Anguilla', 'Switzerland',
       'Sweden', 'British Virgin Islands', 'Czech Republic', 'Mexico',
       'Germany', 'Russia'], dtype=object)
```

```
df['country'].isnull().sum()
```

```
np.int64(62)
```

Finding zero value as input in the data

```
columns = list(df)
columns
```

```
['rank',
 'channel_info',
 'influence_score',
 'posts',
 'followers',
 'avg_likes',
 '60_day_eng_rate',
 'new_post_avg_like',
 'total_likes',
 'country']
```

```
(df[columns[0:9]]==0).sum()
```

	0
rank	0
channel_info	0
influence_score	0
posts	0
followers	0
avg_likes	0
60_day_eng_rate	0
new_post_avg_like	0
total_likes	0

dtype: int64

finding mode for the imputation of the null values in the country column

```
mode = df['country'].mode()
mode
```

	country
0	United States

dtype: object

replacing null values with the mode of the column

```
df['country'].fillna((df['country'].mode()[0]), inplace = True)
```

checking null values again

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

	0
rank	0
channel_info	0
influence_score	0
posts	0
followers	0
avg_likes	0
60_day_eng_rate	0
new_post_avg_like	0
total_likes	0
country	0

dtype: int64

As we get our data clean. Now, converting data into csv file to make dashboards

```
df.to_csv('Cleaned_Insta_Data.csv', index = False)
```

This is a complete process of cleaning a Categorical Data.

THANK YOU!