

Task1_cleaning

September 22, 2025

```
[1]: import pandas as pd
```

```
[2]: df = pd.read_csv('Mall_Customers.csv')
```

```
[3]: df.head()
```

```
[3]:
```

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

```
[4]: df.isnull().sum()
```

```
[4]:
```

CustomerID	0
Gender	0
Age	0
Annual Income (k\$)	0
Spending Score (1-100)	0

dtype: int64

```
[5]: df = df.dropna()
```

```
[6]: df = df.drop_duplicates()
```

```
[7]: df['Gender'] = df['Gender'].str.strip().str.lower()
```

```
[8]: df.columns = [col.strip().lower().replace(' ', '_') for col in df.columns]
```

```
[9]: df['age'] = df['age'].astype(int)
df['annual_income_(k$)'] = pd.to_numeric(df['annual_income_(k$)'],
    ↪errors='coerce')
df['spending_score_(1-100)'] = pd.to_numeric(df['spending_score_(1-100)'],
    ↪errors='coerce')
```

```
[10]: df.head()
```

```
[10]:
```

	customerid	gender	age	annual_income_(k\$)	spending_score_(1-100)
0	1	male	19	15	39
1	2	male	21	15	81
2	3	female	20	16	6
3	4	female	23	16	77
4	5	female	31	17	40

```
[11]: df.to_csv('Mall_Customers_Cleaned.csv', index=False)
```

```
[12]: print("\nSummary of Cleaning:")
print(f"Rows after cleaning: {len(df)}")
print("Columns renamed:", df.columns.tolist())
```

```
Summary of Cleaning:
Rows after cleaning: 200
Columns renamed: ['customerid', 'gender', 'age', 'annual_income_(k$)',
'spending_score_(1-100)']
```

```
[13]: df
```

```
[13]:
```

	customerid	gender	age	annual_income_(k\$)	spending_score_(1-100)
0	1	male	19	15	39
1	2	male	21	15	81
2	3	female	20	16	6
3	4	female	23	16	77
4	5	female	31	17	40
..
195	196	female	35	120	79
196	197	female	45	126	28
197	198	male	32	126	74
198	199	male	32	137	18
199	200	male	30	137	83

```
[200 rows x 5 columns]
```

```
[ ]:
```