```
import pandas as pd
import numpy as np
```

```
# 1. Load dataset
df=pd.read_csv("/Titanic-Dataset.csv")
df.head(5)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S

Next steps: Generate code with df | New interactive sheet

```
df.shape
```

```
(891, 12)
```

```
# 2. Understand data
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

```
# 3. Identify missing values
df.isnull().sum()
```

	0
PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	0
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	0
Embarked	0

dtype: int64

```
# 4. Handle missing values
df.dropna(inplace=True)
```

```
# 4. Handle missing values
df['Age'] = df['Age'].fillna(df['Age'].median())
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
df.drop(columns=['Cabin'], inplace=True)
```

```
# 5. Remove duplicates
df.drop_duplicates(inplace=True)
```

```
# 6.1 Convert datatypes
df['Survived'] = df['Survived'].astype('category')
df['Survived']
```

	Survived
0	0
1	1
2	1
3	1
4	0
...	...
886	0
887	1
888	0
889	1
890	0

891 rows × 1 columns

dtype: category

```
# 6.2 Convert datatypes
df['Pclass'] = df['Pclass'].astype('category')
df['Pclass']
```

	Pclass
0	3
1	1
2	3
3	1
4	3
...	...
886	2
887	1
888	3
889	1
890	3

891 rows × 1 columns

dtype: category

```
# 7.1 Feature engineering
df['Age_Group'] = pd.cut(
    df['Age'],
    bins=[0,12,20,40,60,100],
    labels=['Child','Teen','Adult','Middle_Aged','Senior']
)
df['Age_Group']
```

	Age_Group
0	Adult
1	Adult
2	Adult
3	Adult
4	Adult
...	...
886	Adult
887	Teen
888	NaN
889	Adult
890	Adult

891 rows × 1 columns

dtype: category

```
# 7.2 Feature engineering
df['Fare_Band'] = pd.qcut(df['Fare'], 4, labels=['Low','Medium','High','Very High'])
df['Fare_Band']
```

	Fare_Band
0	Low
1	Very High
2	Medium
3	Very High
4	Medium
...	...
886	Medium
887	High
888	High
889	High
890	Low

891 rows × 1 columns

dtype: category

```
# 8. Save cleaned dataset
df.to_csv("Titanic_Cleaned.csv", index=False)

print("Cleaned file saved successfully!")
```

```
Cleaned file saved successfully!
```