Titanic Missing Values Handling

Source code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# Load dataset
df = pd.read_csv("titanic.csv")
# Display basic information
df.info()
print(df.head())
# Check for missing values
missing_values = df.isnull().sum()
print("Missing Values:\n", missing_values[missing_values > 0])
# Visualizing missing data
plt.figure(figsize=(10,6))
sns.heatmap(df.isnull(), cmap="viridis", cbar=False, yticklabels=False)
plt.title("Missing Values Heatmap")
plt.show()
# Handling Missing Values
## Age: Fill with median
df["Age"].fillna(df["Age"].median(), inplace=True)
## Cabin: Fill with 'Unknown'
df["Cabin"].fillna("Unknown", inplace=True)
## Embarked: Fill with mode
df["Embarked"].fillna(df["Embarked"].mode()[0], inplace=True)
## Fare: Fill with median
df["Fare"].fillna(df["Fare"].median(), inplace=True)
# Handling specific row data manually
```

```
df.loc[df['Name'] == 'Sivola, Mr. Antti Wilhelm', 'Cabin'] = 'Unknown'
# Verify missing values are handled
print("After Imputation:\n", df.isnull().sum())
# Display the specific processed row
specific_row = df[df['Name'] == 'Sivola, Mr. Antti Wilhelm']
print("Processed Row:\n", specific_row)
# Save cleaned dataset
df.to_csv("titanic_cleaned.csv", index=False)
print("titanic_cleaned.csv saved successfully.")
```

The output of the program is:

1. Basic Information (`df.info()`)

This part of the program shows details about the dataset, including:

- How many rows and columns it has (891 rows and 12 columns).
- Which columns have missing values.
- The type of data in each column (numbers, text, etc.).

Example output:

- The **Age** column has **177 missing values**.
- The **Cabin** column has **687 missing values**.
- The **Embarked** column has **2 missing values**.
- The **Fare** column may have some missing values in certain datasets.

---

2. First 5 Rows (`df.head()`)

This part displays the first five rows of the dataset. It helps us see how the data looks.

Example:

PassengerId	Survived	Pclass	Name	Sex	Age	Cabin	Fare	Embarked
1	0	3	**Some Passenger Name 1**	**male**	22	**NaN**	7.25	S
2	1	1	**Some Passenger Name 2**	**female**	38	C85	71.28	C
3	1	3	**Some Passenger Name 3**	**female**	26	**NaN**	7.92	S
4	1	1	**Some Passenger Name 4**	**male**	35	C123	53.1	S
5	0	3	**Some Passenger Name 5**	**male**	35	**NaN**	8.05	S

From this, we can see that some values in the **Cabin, Age, and Embarked** columns are missing.

3. Checking Missing Values

Before fixing missing values, the program counts how many are missing.

Example output:

```
Missing Values:
Age       → 177 missing
Cabin     → 687 missing
Embarked  → 2 missing
Fare      → 1 missing
```

4. Missing Values Heatmap (`sns.heatmap(df.isnull())`)

The program creates a heatmap (a visual chart) to show where values are missing.

- Missing values appear as bright spots.
- This helps us see which columns have the most missing data.

5. After Fixing Missing Values

The program fills missing values using different methods:

- **Age** → Replaced with the **median age**.
- **Cabin** → Replaced with **"Unknown"**.
- **Embarked** → Replaced with the most common value (**mode**).
- **Fare** → Replaced with the **median fare**.

After fixing these, the program checks again, and now there are **no missing values**!

Example output:

```
After Imputation (fixing missing values):

PassengerId  → 0 missing

Survived     → 0 missing

Pclass       → 0 missing

Age          → 0 missing

Cabin        → 0 missing

Fare         → 0 missing

Embarked     → 0 missing
```

This confirms all missing values have been handled.

6. Updating a Specific Row (Sivola, Mr. Antti Wilhelm)

The program also manually updates the **Cabin** value for a passenger named **"Sivola, Mr. Antti Wilhelm"**, setting it to `"Unknown"`.

Example:

```
Processed Row:

Passenger Name → Sivola, Mr. Antti Wilhelm

Age → 21

Cabin → Unknown

Fare → 7.925

Embarked → S
```

This confirms that the change was successfully applied.

7. Saving the Cleaned Dataset

Finally, the cleaned data is saved to a new file called **"titanic_cleaned.csv"**, which now has no missing values.