HW1 Python

September 12, 2024

[1]:

```
import pandas as pd

# Load the dataset
url = 'https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.
 ↪csv'
df = pd.read_csv(url)

# Check for missing values
missing_values = df.isnull().sum()

# Display the count of missing values
print(missing_values)
```

```
Survived                   0
Pclass                     0
Name                       0
Sex                        0
Age                        0
Siblings/Spouses Aboard    0
Parents/Children Aboard    0
Fare                       0
dtype: int64
```

[2]:

```
import pandas as pd
url = "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
 ↪data/2020/2020-05-05/villagers.csv"
df = pd.read_csv(url)

# Check for missing values
missing_values = df.isna().sum()

# Display the count of missing values in each column
print(missing_values)
```

```
row_n          0
id             1
name           0
gender         0
```

```
species         0
birthday        0
personality     0
song           11
phrase          0
full_id         0
url             0
dtype: int64
```

[3]:

```python
import pandas as pd
url = "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
 ↪data/2020/2020-05-05/villagers.csv"
df = pd.read_csv(url)
print(df.isna().sum())
```

```
row_n           0
id              1
name            0
gender          0
species         0
birthday        0
personality     0
song           11
phrase          0
full_id         0
url             0
dtype: int64
```

[7]:

```python
import pandas as pd
url = 'https://raw.githubusercontent.com/jbrownlee/Datasets/master/housing.csv'
house_data = pd.read_csv(url, header=None)
print(house_data.isnull().sum())
```

```
0     0
1     0
2     0
3     0
4     0
5     0
6     0
7     0
8     0
9     0
10    0
11    0
12    0
13    0
dtype: int64
```

[5]:

```
import pandas as pd

# Load the dataset
url = "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
 ↪data/2020/2020-05-05/villagers.csv"
df = pd.read_csv(url)

# Check for missing values
missing_values = df.isna().sum()

# Print missing values
print("Missing Values in Each Column:\n", missing_values)

# Summary of the dataset: basic statistics for numerical columns and an␣
 ↪overview for categorical columns
summary = df.describe(include='all')

# Print the summary
print("\nSummary of the Dataset:\n", summary)
```

```
Missing Values in Each Column:
 row_n           0
id             1
name           0
gender         0
species        0
birthday       0
personality    0
song          11
phrase         0
full_id        0
url            0
dtype: int64

Summary of the Dataset:
              row_n       id     name gender species birthday personality  \
count   391.000000      390      391    391     391      391         391
unique         NaN      390      391      2      35      361           8
top            NaN  admiral  Admiral   male     cat     1-27        lazy
freq           NaN        1        1    204      23        2          60
mean    239.902813      NaN      NaN    NaN     NaN      NaN         NaN
std     140.702672      NaN      NaN    NaN     NaN      NaN         NaN
min       2.000000      NaN      NaN    NaN     NaN      NaN         NaN
25%     117.500000      NaN      NaN    NaN     NaN      NaN         NaN
50%     240.000000      NaN      NaN    NaN     NaN      NaN         NaN
75%     363.500000      NaN      NaN    NaN     NaN      NaN         NaN
max     483.000000      NaN      NaN    NaN     NaN      NaN         NaN
```

```
                song   phrase           full_id  \
count            380      391               391
unique            92      388               391
top     K.K. Country  wee one  villager-admiral
freq              10        2                 1
mean             NaN      NaN               NaN
std              NaN      NaN               NaN
min              NaN      NaN               NaN
25%              NaN      NaN               NaN
50%              NaN      NaN               NaN
75%              NaN      NaN               NaN
max              NaN      NaN               NaN

                                                      url
count                                                 391
unique                                                391
top     https://villagerdb.com/images/villagers/thumb/…
freq                                                    1
mean                                                  NaN
std                                                   NaN
min                                                   NaN
25%                                                   NaN
50%                                                   NaN
75%                                                   NaN
max                                                   NaN
```

[8]:

```
import pandas as pd

# Load your dataset
url = "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
↪data/2020/2020-05-05/villagers.csv"
df = pd.read_csv(url)

# Get the shape of the dataset
print("Shape of the dataset (rows, columns):", df.shape)

# Describe the dataset (numeric columns only)
summary = df.describe()

# Print summary
print("\nSummary statistics for numeric columns:\n", summary)

# Check for missing values
print("\nMissing values in each column:\n", df.isna().sum())
```

```
Shape of the dataset (rows, columns): (391, 11)
```

```
Summary statistics for numeric columns:
             row_n
count  391.000000
mean   239.902813
std    140.702672
min      2.000000
25%    117.500000
50%    240.000000
75%    363.500000
max    483.000000

Missing values in each column:
 row_n           0
id             1
name           0
gender         0
species        0
birthday       0
personality    0
song          11
phrase         0
full_id        0
url            0
dtype: int64
```

```
[ ]:
```