Pre-lecture week1 HW

September 12, 2024

[1]:
```
import pandas as pd
url = "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
 ↪data/2020/2020-05-05/villagers.csv"
df = pd.read_csv(url)
df.isna().sum()
```

[1]:
```
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]:
```
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)

# Get the number of rows and columns
num_rows, num_cols = df.shape
print(f"Number of rows: {num_rows}")
print(f"Number of columns: {num_cols}")

# Display column names
print("\nColumn names:")
print(df.columns.tolist())

# Check for missing data in each column
missing_data = df.isna().sum()
```

```
print("\nMissing data in each column:")
print(missing_data)
```

```
Number of rows: 391
Number of columns: 11

Column names:
['row_n', 'id', 'name', 'gender', 'species', 'birthday', 'personality', 'song',
'phrase', 'full_id', 'url']

Missing data 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
```

[4]:
```
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)

# Summary for numerical columns
print("Summary for numerical columns:")
print(df.describe(include=[float, int]))

# Summary for categorical columns
print("\nSummary for categorical columns:")
print(df.describe(include=[object]))

# Summary for all columns
print("\nSummary for all columns:")
print(df.info())
```

```
Summary for numerical 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

Summary for categorical columns:
              id     name gender species birthday personality          song  \
count         390      391    391     391      391         391           380
unique        390      391      2      35      361           8            92
top       admiral  Admiral   male     cat     1-27        lazy  K.K. Country
freq            1        1    204      23        2          60            10

          phrase           full_id  \
count        391               391
unique       388               391
top     wee one  villager-admiral
freq           2                 1

                                                      url
count                                                 391
unique                                                391
top     https://villagerdb.com/images/villagers/thumb/…
freq                                                    1

Summary for all columns:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 391 entries, 0 to 390
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   row_n        391 non-null    int64
 1   id           390 non-null    object
 2   name         391 non-null    object
 3   gender       391 non-null    object
 4   species      391 non-null    object
 5   birthday     391 non-null    object
 6   personality  391 non-null    object
 7   song         380 non-null    object
 8   phrase       391 non-null    object
 9   full_id      391 non-null    object
 10  url          391 non-null    object
dtypes: int64(1), object(10)
memory usage: 33.7+ KB
None
```

[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)

# Identify non-numeric variables
non_numeric_columns = df.select_dtypes(exclude=['number']).columns
print("Non-numeric columns:")
print(non_numeric_columns)

# Identify missing values in numeric variables
numeric_columns = df.select_dtypes(include=['number']).columns
missing_values_numeric = df[numeric_columns].isna().sum()
print("\nMissing values in numeric columns:")
print(missing_values_numeric)
```

```
Non-numeric columns:
Index(['id', 'name', 'gender', 'species', 'birthday', 'personality', 'song',
       'phrase', 'full_id', 'url'],
      dtype='object')

Missing values in numeric columns:
row_n    0
dtype: int64
```

[ ]: