Data Summarization HW due Sept 12

September 12, 2024

[1]:
```
import pandas as pd

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

# Check for missing values in the dataset
missing_values = df.isna().sum()
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
```

[2]:
```
import pandas as pd

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

# Check for missing values (NA) in the dataframe
df.isna().sum()
```

[2]:
```
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

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

# Generate summary statistics for the dataset (only numerical columns)
print("Numerical Data Summary:")
print(df.describe())

# Generate a summary of categorical columns
print("\nCategorical Data Summary:")
categorical_columns = df.select_dtypes(include=['object']).columns
for col in categorical_columns:
    print(f"\nSummary of '{col}' column:")
    print(df[col].value_counts())
```

```
Numerical Data Summary:
            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

Categorical Data Summary:

Summary of 'id' column:
id
admiral    1
mott       1
paula      1
patty      1
```

```
pate       1
          ..
eloise     1
elmer      1
ellie      1
elise      1
zucker     1
Name: count, Length: 390, dtype: int64

Summary of 'name' column:
name
Admiral    1
Muffy      1
Paula      1
Patty      1
Pate       1
          ..
Elvis      1
Eloise     1
Elmer      1
Ellie      1
Zucker     1
Name: count, Length: 391, dtype: int64

Summary of 'gender' column:
gender
male      204
female    187
Name: count, dtype: int64

Summary of 'species' column:
species
cat          23
rabbit       20
frog         18
squirrel     18
duck         17
dog          16
cub          16
pig          15
bear         15
mouse        15
horse        15
bird         13
penguin      13
sheep        13
elephant     11
wolf         11
```

```
ostrich      10
deer         10
eagle         9
gorilla       9
chicken       9
koala         9
goat          8
hamster       8
kangaroo      8
monkey        8
anteater      7
hippo         7
tiger         7
alligator     7
lion          7
bull          6
rhino         6
cow           4
octopus       3
Name: count, dtype: int64

Summary of 'birthday' column:
birthday
1-27     2
12-5     2
7-31     2
3-26     2
8-3      2
        ..
4-3      1
10-26    1
7-23     1
12-8     1
3-8      1
Name: count, Length: 361, dtype: int64

Summary of 'personality' column:
personality
lazy      60
normal    59
cranky    55
snooty    55
jock      55
peppy     49
smug      34
uchi      24
Name: count, dtype: int64
```

```
Summary of 'song' column:
song
K.K. Country     10
Forest Life       9
Imperial K.K.     7
K.K. Soul         7
K.K. Ragtime      7
                 ..
Aloha K.K.        2
Drivin'           1
Senor K.K.        1
K.K.  Bazaar      1
K.K. D&B          1
Name: count, Length: 92, dtype: int64

Summary of 'phrase' column:
phrase
wee one       2
quacko        2
bloop         2
aye aye       1
snoot         1
              ..
lambchop      1
yeah buddy    1
chow down     1
unh-hunh      1
pronk         1
Name: count, Length: 388, dtype: int64

Summary of 'full_id' column:
full_id
villager-admiral    1
villager-muffy      1
villager-paula      1
villager-patty      1
villager-pate       1
                   ..
villager-elvis      1
villager-eloise     1
villager-elmer      1
villager-ellie      1
villager-zucker     1
Name: count, Length: 391, dtype: int64

Summary of 'url' column:
url
https://villagerdb.com/images/villagers/thumb/admiral.98206ee.png    1
```

```
https://villagerdb.com/images/villagers/thumb/muffy.1497c92.png      1
https://villagerdb.com/images/villagers/thumb/paula.563ba81.png      1
https://villagerdb.com/images/villagers/thumb/patty.3e17f7f.png      1
https://villagerdb.com/images/villagers/thumb/pate.c60838c.png       1
                                                                     ..
https://villagerdb.com/images/villagers/thumb/elvis.57d4757.png      1
https://villagerdb.com/images/villagers/thumb/eloise.112208b.png     1
https://villagerdb.com/images/villagers/thumb/elmer.cc7df52.png      1
https://villagerdb.com/images/villagers/thumb/ellie.5a144a6.png      1
https://villagerdb.com/images/villagers/thumb/zucker.8dbb719.png     1
Name: count, Length: 391, dtype: int64
```

[4]:
```python
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 shape of the dataset
print("Shape of the dataset (rows, columns):", df.shape)

# Display summary statistics for numeric columns
print("\nSummary statistics for numeric columns:")
print(df.describe())

# Display count of non-missing values in each column (numeric and non-numeric)
print("\nCount of non-missing values in each column:")
print(df.count())

# Check for missing values in numeric columns
print("\nNumber of missing values in each numeric column:")
numeric_columns = df.select_dtypes(include=['number']).columns
print(df[numeric_columns].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
```

```
Count of non-missing values in each column:
row_n          391
id             390
name           391
gender         391
species        391
birthday       391
personality    391
song           380
phrase         391
full_id        391
url            391
dtype: int64

Number of missing values in each numeric column:
row_n    0
dtype: int64
```

[5]:
```python
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)

# Attribute example: df.shape
# This gives the shape of the DataFrame as a tuple (rows, columns)
print("Attribute - df.shape:")
print(df.shape)

# Method example: df.describe()
# This calculates and returns summary statistics for numeric columns
print("\nMethod - df.describe():")
print(df.describe())

# Additional method example: df.head() - This shows the first few rows of the␣
 ↪DataFrame
print("\nMethod - df.head():")
print(df.head())
```

```
Attribute - df.shape:
(391, 11)

Method - df.describe():
            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

Method - df.head():
   row_n       id     name  gender    species birthday personality  \
0      2  admiral  Admiral    male       bird     1-27      cranky
1      3  agent-s  Agent S  female   squirrel      7-2       peppy
2      4    agnes    Agnes  female        pig     4-21        uchi
3      6       al       Al    male    gorilla    10-18        lazy
4      7  alfonso  Alfonso    male  alligator      6-9        lazy

          song    phrase           full_id  \
0   Steep Hill   aye aye  villager-admiral
1      DJ K.K.  sidekick  villager-agent-s
2   K.K. House   snuffle    villager-agnes
3   Steep Hill   Ayyeeee       villager-al
4  Forest Life  it'sa me  villager-alfonso

                                                 url
0  https://villagerdb.com/images/villagers/thumb/…
1  https://villagerdb.com/images/villagers/thumb/…
2  https://villagerdb.com/images/villagers/thumb/…
3  https://villagerdb.com/images/villagers/thumb/…
4  https://villagerdb.com/images/villagers/thumb/…
```

[ ]: