01 December 2022

UNIVERSITY OF GLOUCESTERSHIRE

CT4031
Maths for Data Science

Week 3 – Practical

Data cleaning

- Detecting missing values
- Dropping columns
- Changing indexes
- Tidying up Fields in the Data
- Renaming Columns and Skipping Rows

Detecting missing values - 1

```
# Importing libraries
import pandas as pd
import numpy as np

# Read csv file into a pandas dataframe
df = pd.read_csv("olympics.csv")

# Take a look at the first few rows
print (df.head())

# searching for missing values
print (df['0'].isnull())
```

Detecting missing values - 2

```
# Importing libraries
import pandas as pd
import numpy as np

# Making a list of missing value types
missing_values = ["n/a", "na", "--", "??"]
df = pd.read_csv("olympics.csv", na_values = missing_values)

# searching for missing values
print (df['0'].isnull())
```

Dropping columns

```
# Importing libraries
import pandas as pd
import numpy as np

# Read csv file into a pandas dataframe
df = pd.read_csv("olympics.csv")

to_drop = ['0', '1']

df.drop(to_drop, inplace=True, axis=1)

print(df.head())
```

Changing indexes

```
# Importing libraries
import pandas as pd
import numpy as np

# Read csv file into a pandas dataframe
df = pd.read_csv("olympics.csv")

print(df['0'].is_unique)

#changing the index
df = df.set_index('0')

print(df.head())
```

Tidying up Fields in the Data

```
# Importing libraries
import pandas as pd
import numpy as np

# Read csv file into a pandas dataframe
df = pd.read_csv("olympics.csv")

# Extracting the code
extr = df['0'].str.extract(r'^(\d{1})', expand=False)
print(extr.head())

# Converting to numeric
df['0'] = pd.to_numeric(extr)
print(df.head())
```

Renaming Columns and Skipping Rows

```
import pandas as pd
import numpy as np
# Read csv file into a pandas dataframe
df = pd.read_csv("olympics.csv", header=1)
print (df.head())
#changing the names
new_names =  {'Unnamed: 0': 'Country','? Summer': 'Summer Olympics',
              '01 !': 'Gold','02 !': 'Silver',
              '03 !': 'Bronze','? Winter': 'Winter Olympics',
              '01 !.1': 'Gold.1','02 !.1': 'Silver.1',
              '03 !.1': 'Bronze.1','? Games': '# Games',
              '01 !.2': 'Gold.2','02 !.2': 'Silver.2',
              '03 !.2': 'Bronze.2'}
df.rename(columns=new_names, inplace=True)
print (df.head())
```

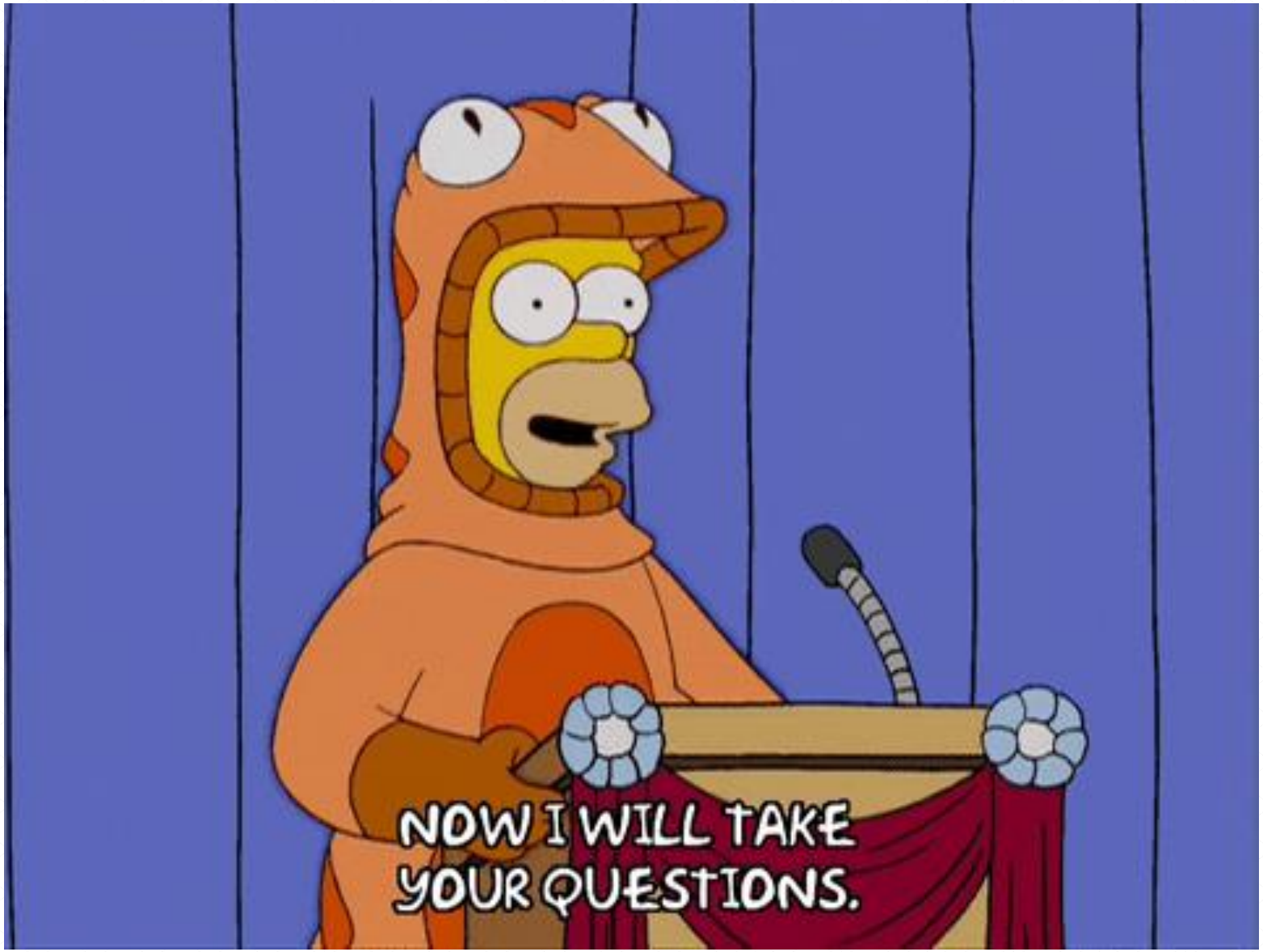
Practice

Using the *books* dataset, practice the concepts learnt during today's lesson.

References

Data Cleaning with Python and Pandas: Detecting Missing Values | by John Sullivan | Towards Data Science [WWW Document], n.d. URL https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b (accessed 2.22.21).

Python, R., n.d. Pythonic Data Cleaning With Pandas and NumPy – Real Python [WWW Document]. URL https://realpython.com/python-data-cleaning-numpy-pandas/ (accessed 2.22.21).


NOW I WILL TAKE
YOUR QUESTIONS.