Undertanding the cartwheel data set

The notebook aims to undertand the content of the cartwheel data set.

Acknowledgments

- Data from https://www.coursera.org/ from the course "Understanding and Visualizing Data with Python" by University of Michigan

Cartwheel data set

1. A cartwheel

cartwheel1.png

2. The dataset description
    - The dataset used here is an extension from the original cartwheel dataset from cursera
    - Total numer of observations: 52
    - Many observations/measurements/recordings of the characteristics/attributes/ variables of cartwheel executions
    - Variables: Age, Gender, GenderGroup, Glasses, GlassesGroup, Height, Wingspan, CWDistance, ... (X variables)

Importing and inspecting the data

```
# Define where you are running the code: colab or local
RunInColab          = True     # (False: no  | True: yes)

# If running in colab:
if RunInColab:
    # Mount your google drive in google colab
    from google.colab import drive
    drive.mount('/content/drive')

    # Find location
    #!pwd
    #!ls
```

```
    #!ls
    #!ls "/content/drive/My Drive/TC1002S/NotebooksProfessor"

    # Define path del proyecto
    Ruta            = "/content/drive/My Drive/TC1002S/NotebooksProfessor"

else:
    # Define path del proyecto
    Ruta            = ""
```

```
Mounted at /content/drive
```

```
# Import the packages that we will be using
import matplotlib.pyplot as plt
import pandas as pd

# Dataset url
url = Ruta + "/datasets/cartwheel/cartwheel.csv"

# Load the dataset
dataset = pd.read_csv(url )


# Print the dataset
dataset
```

	ID	Age	Gender	GenderGroup	Glasses	GlassesGroup	Height	Wingspan	CWDistance
0	1	56.0	F	1	Y	1	62.00	61.0	79
1	2	26.0	F	1	Y	1	62.00	60.0	70
2	3	33.0	F	1	Y	1	66.00	64.0	85
3	4	39.0	F	1	N	0	64.00	63.0	87
4	5	27.0	M	2	N	0	73.00	75.0	72
5	6	24.0	M	2	N	0	75.00	71.0	81
6	7	28.0	M	2	N	0	75.00	76.0	107
7	8	22.0	F	1	N	0	65.00	62.0	98
8	9	29.0	M	2	Y	1	74.00	73.0	106
9	10	33.0	F	1	Y	1	63.00	60.0	65
10	11	30.0	M	2	Y	1	69.50	66.0	96
11	12	28.0	F	1	Y	1	62.75	58.0	79
12	13	25.0	F	1	Y	1	65.00	64.5	92
13	14	23.0	F	1	N	0	61.50	57.5	66

14	15	31.0	M	2	Y	1	73.00	74.0	72
15	16	26.0	M	2	Y	1	71.00	72.0	115
16	17	26.0	F	1	N	0	61.50	59.5	90
17	18	27.0	M	2	N	0	66.00	66.0	74
18	19	23.0	M	2	Y	1	70.00	69.0	64
19	20	24.0	F	1	Y	1	68.00	66.0	85
20	21	23.0	M	2	Y	1	69.00	67.0	66
21	22	29.0	M	2	N	0	71.00	70.0	101
22	23	25.0	M	2	N	0	70.00	68.0	82
23	24	26.0	M	2	N	0	69.00	71.0	63
24	25	23.0	F	1	Y	1	65.00	63.0	67
25	26	28.0	M	2	N	0	75.00	76.0	111
26	27	24.0	M	2	N	0	78.40	71.0	92
27	28	25.0	M	2	Y	1	76.00	73.0	107
28	29	32.0	F	1	Y	1	63.00	60.0	75
29	30	38.0	F	1	Y	1	61.50	61.0	78
30	31	27.0	F	1	Y	1	62.00	60.0	72

```
# Print the number of rows
Nrows = dataset.shape[0]
Nrows
```

```
52
```

34	35	24.0	F	1	N	0	67.80	62.0	98

```
# Print the number of columns
Ncols = dataset.shape[1]
Ncols
```

```
12
```

Data types

```
dataset.dtypes
```

```
ID                int64
Age             float64
Gender           object
GenderGroup       int64
Glasses          object
```

```
GlassesGroup       int64
Height           float64
Wingspan         float64
CWDistance         int64
Complete          object
CompleteGroup    float64
Score              int64
dtype: object
```

Activity: work with the iris dataset

1. Load the iris.csv file in your computer and understand the dataset
2. How many observations (rows) are in total?
3. How many variables (columns) are in total? What do they represent?
4. How many observations are for each type of flower?
5. What is the type of data for each variable?
6. What are the units of each variable?

1. Loading the Iris dataset

```
# Dataset url
url = Ruta + "/datasets/iris/iris.csv"

# Load the dataset
newHeader=["Sepal_length", "Sepal_width", "Petal_length", "Petal_width", "Class"]
dataset2 = pd.read_csv(url, header=None, names=newHeader )

#dataset = dataset.rename(columns={"Sepal_length": 0, "Sepal_width": 1, "Petal_length": 2

# Print the dataset
dataset2
```

	Sepal_length	Sepal_width	Petal_length	Petal_width	Class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa

4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

2. Total rows

```
# Print the number of rows
Nrows = dataset2.shape[0]
Nrows
```

```
150
```

3. Total columns

The columns represent the variables collected in each instance, with each column being a different variable

```
# Print the number of columns
Ncols = dataset2.shape[1]
Ncols
```

```
5
```

4. Total observations for each type of flower

```
print("Instances of Iris-setosa: ", dataset2['Class'].value_counts()['Iris-setosa'], "\n"

print("Instances of Iris-versicolor: ", dataset2['Class'].value_counts()['Iris-versicolor

print("Instances of Iris-virginica: ", dataset2['Class'].value_counts()['Iris-virginica']
```

```
Instances of Iris-setosa:  50
```

```
Instances of Iris-versicolor:  50

Instances of Iris-virginica:  50
```

5. Type of data for each variable

```
dataset2.dtypes
```

```
Sepal_length    float64
Sepal_width     float64
Petal_length    float64
Petal_width     float64
Class            object
dtype: object
```

6. Units for each variable

As specified on the .names file, the first four columns are measurements of each flower observed, and they're measured in cm. As for the fifth column, it refers to the name of the class of the flower observed in that row.

```
dataset2.head()
```

	Sepal_length	Sepal_width	Petal_length	Petal_width	Class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa