Multivariate Analysis

November 3, 2022

In this assignment we'll ask you to plot multiple variables.

You will use what you find in this assignment to answer the questions in the quiz that follows. It may be useful to keep this notebook side-by-side with this week's quiz on your screen.

```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import scipy.stats as stats
        %matplotlib inline
        import matplotlib.pyplot as plt
        pd.set_option('display.max_columns', 100)

        path = "Cartwheeldata.csv"
```

```
In [2]: # First, you must import the cartwheel data from the path given above
        df = pd.read_csv('Cartwheeldata.csv') # using pandas, read in the csv data found at th
```

```
In [3]: # Next, look at the 'head' of our DataFrame 'df'.
        df.head()
```

```
Out[3]:    ID  Age Gender  GenderGroup Glasses  GlassesGroup  Height  Wingspan  \
        0   1   56      F            1       Y             1    62.0      61.0
        1   2   26      F            1       Y             1    62.0      60.0
        2   3   33      F            1       Y             1    66.0      64.0
        3   4   39      F            1       N             0    64.0      63.0
        4   5   27      M            2       N             0    73.0      75.0

           CWDistance Complete  CompleteGroup  Score
        0          79        Y              1      7
        1          70        Y              1      8
        2          85        Y              1      7
        3          87        Y              1     10
        4          72        N              0      4
```

If you can't remember a function, open a previous notebook or video as a reference, or use your favorite search engine to look for a solution.

```
In [4]: df.columns
```

```
Out[4]: Index(['ID', 'Age', 'Gender', 'GenderGroup', 'Glasses', 'GlassesGroup',
               'Height', 'Wingspan', 'CWDistance', 'Complete', 'CompleteGroup',
               'Score'],
               dtype='object')
```

0.1 Scatter plots

First, let's looks at two variables that we expect to have a strong relationship, 'Height' and 'Wingspan'.

```
In [14]: # Make a Seaborn scatter plot with x = height and y = wingspan using sns.scatterplot(
         x = df['Height']
         y = df['Wingspan']
         sns.scatterplot(x,y, hue = df['Gender']);
```
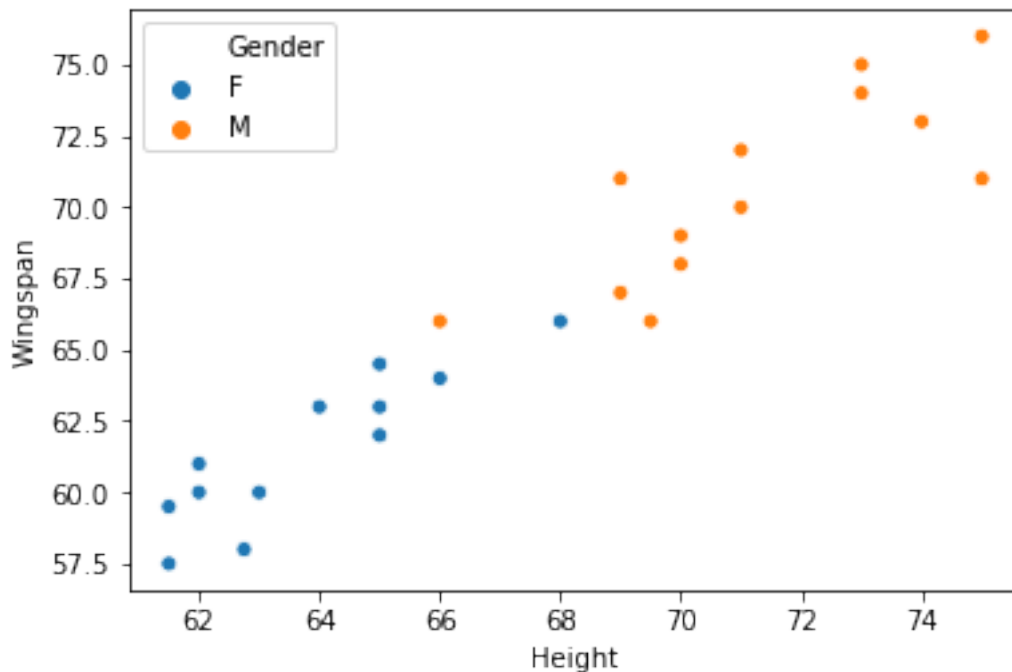

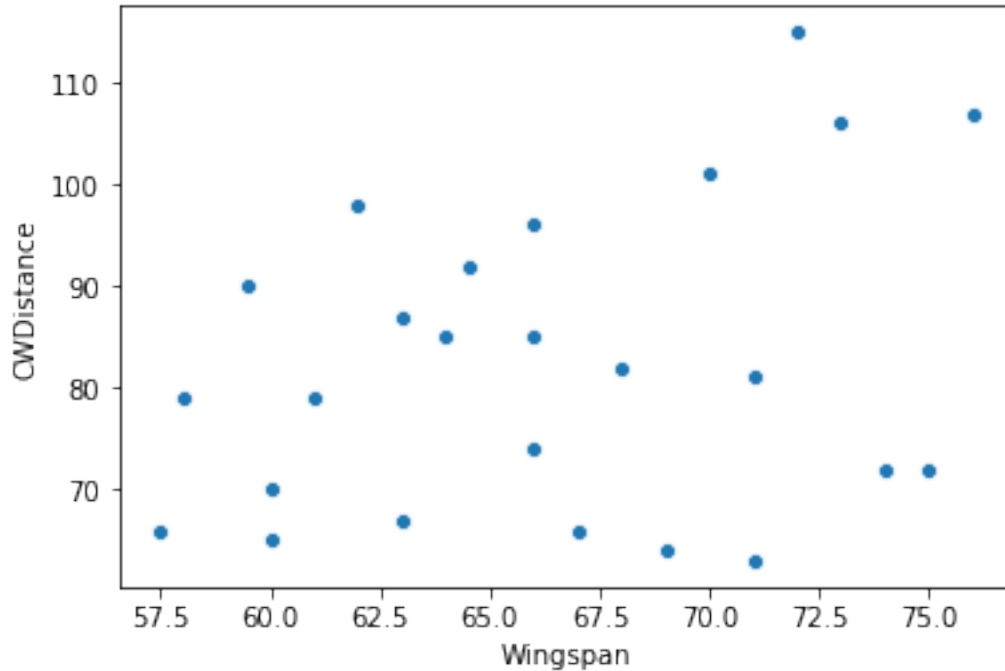How would you describe the relationship between 'Height' and 'Wingspan'?
Questions you can ask: * Is it linear? * Are there outliers? * Are their ranges similar or different?
How else could you describe the relationship?
From the scatter plot we can see that there is a linear relationship between wingspan and height. When breaking down the scatterplot by gender this is constant across gender. There is some variation with outlying points however the clustering of the points do not take away from the trend of the data. The range between male and female data is different.
Now let's look at two variables that we don't yet assume have a strong relationship, 'Wingspan' and 'CWDistance'

2

```
In [8]:  # Make a Seaborn scatter plot with x = wingspan and y = cartwheel distance
         x = df['Wingspan']
         y = df['CWDistance']
         sns.scatterplot(x,y);
```



```
In [16]: df[['CWDistance', 'Wingspan']].describe()

Out[16]:           CWDistance    Wingspan
         count     25.000000   25.000000
         mean      82.480000   66.260000
         std       15.058552    5.492647
         min       63.000000   57.500000
         25%       70.000000   62.000000
         50%       81.000000   66.000000
         75%       92.000000   71.000000
         max      115.000000   76.000000
```

How would you describe the relationship between 'Wingspan' and 'CWDistance'?
* Is it linear? * Are there outliers? * Are their ranges similar or different?

How else could you describe the relationship?

Let makes the same plot as above, but now include 'Gender' as the color scheme by including the argument

```
hue=df['Gender']
```

in the Seaborn function

The data is showing a non-linear relationship. The data does not show a trendline with many outliers present giving the interpretation that there is no direct relationship.

```
In [9]: # Make a Seaborn scatter plot with x = wingspan and y = cartwheel distance, and hue =
        x = df['Wingspan']
        y = df['CWDistance']
        sns.scatterplot(x,y, hue=df['Gender']);
```
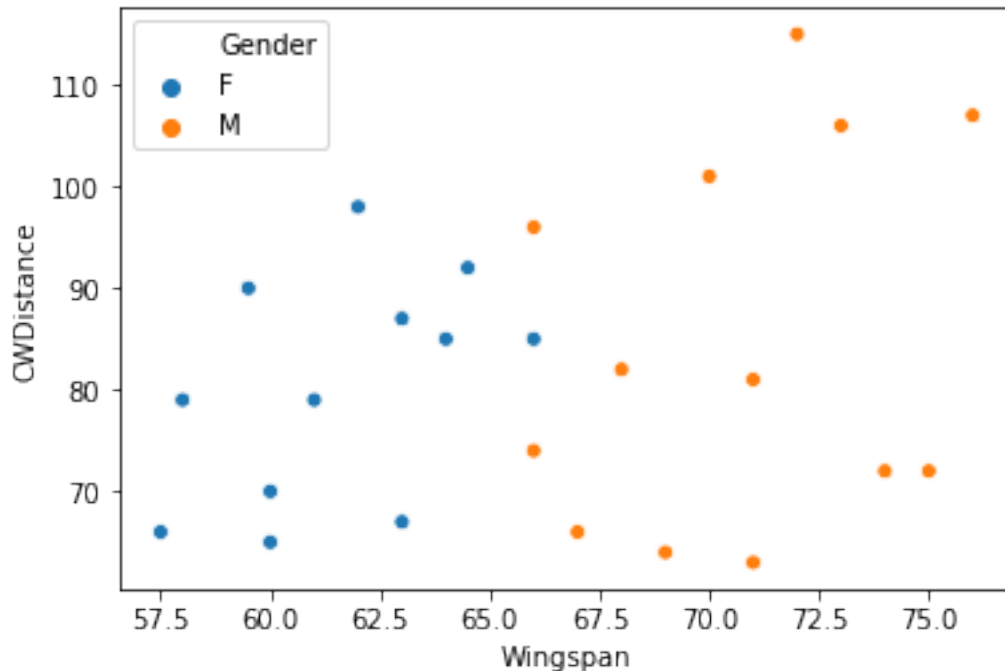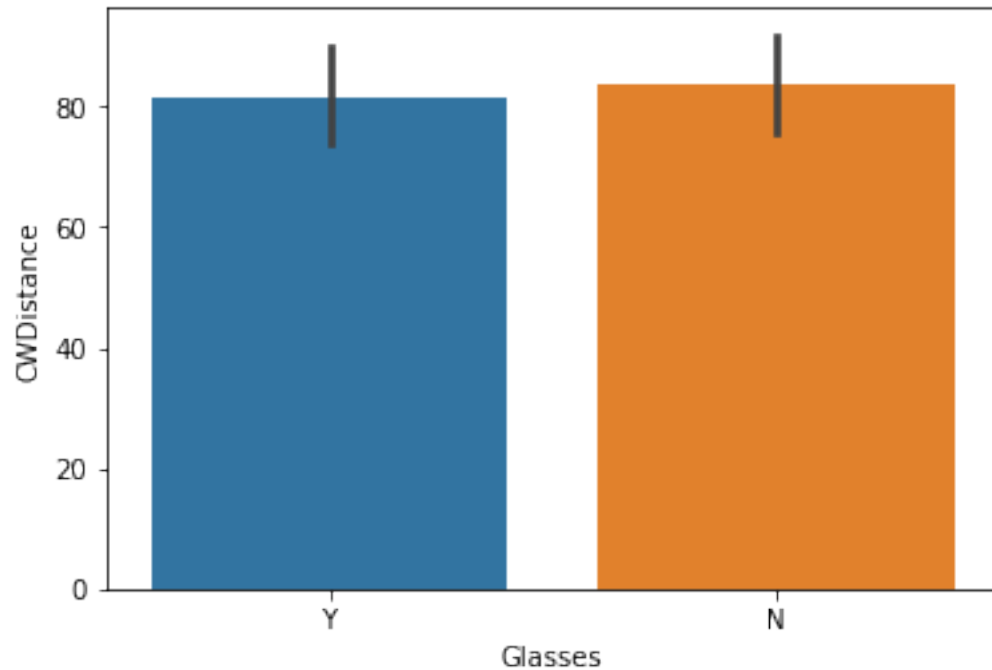

Does does this new information on the plot change your interpretation of the relationship between 'Wingspan' and 'CWDistance'?

0.2 Barcharts

Now lets plot barplots of 'Glasses'

```
In [10]: # Make a Seaborn barplot with x = glasses and y = cartwheel distance
         x = df['Glasses']
         y = df['CWDistance']
         sns.barplot(x,y)
```
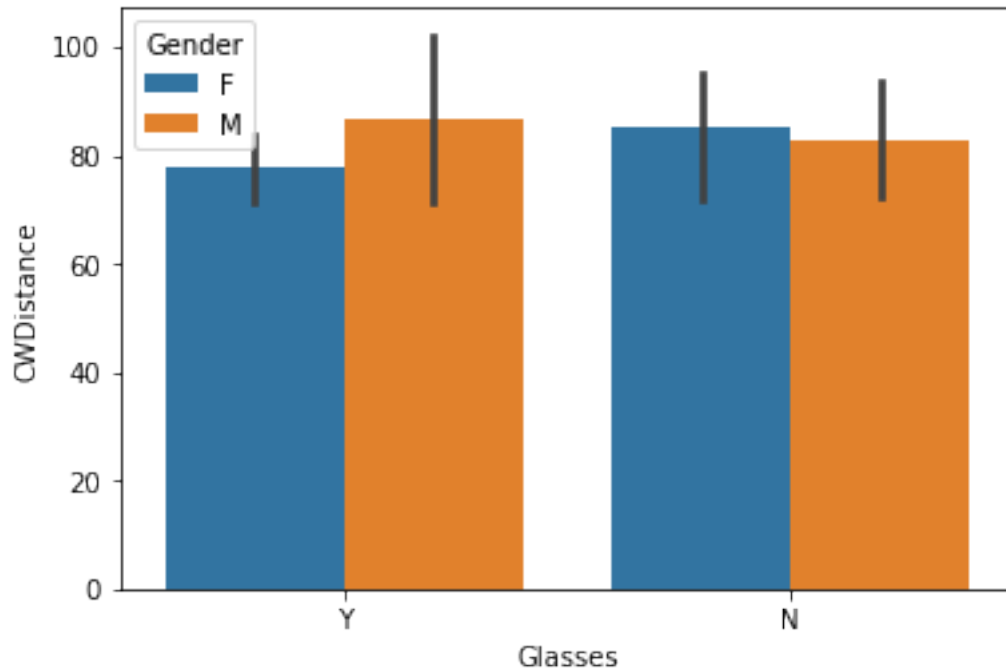
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x7f82e5b9a5f8>

4

What can you say about the relationship of 'Glasses' and 'CWDistance'?

The relationship between those with glasses and their cartwheel distance is that in this boxplot example, those without glasses have a larger cart wheel distance.

In [13]: # Make the same Seaborn boxplot as above, but include gender for the hue argument
         x = df['Glasses']
         y = df['CWDistance']
         sns.barplot(x,y, hue = df['Gender']);

How does this new plot change your interpretation about the relationship of 'Glasses' and 'CWDistance'?

When stratifying the data by gender, the assumption made above is no longer applicable, there is variation between those with glasses and those without alongside gender being a factor.