DA3: selecting columns

We are practicing selecting columns from our data and then creating new columns and then using them for plotting.

In [ ]:
```
import pandas as pd
from plotnine import *
import numpy as np

surveys = pd.read_csv("data/surveys.csv")

surveys.columns
```

Scale the 'weight' column manually creating a new column

We use the **.mean()** and **.std()** methods to calculate mean & standard deviation for scaling.

In [ ]:
```
surveys['weight_scaled'] = (surveys['weight'] - surveys['weight'].mean())
```

NB: In geom_jitter the width parameter controls the horizontal jitter. And in geom_hline we use yintercept to position the horizontal and the color and size parameters to set the style of the line.

In [ ]:
```
p = (
    ggplot(surveys, aes(x='species_id', y='weight_scaled')) +
    geom_jitter(width=0.1) +
    geom_hline(yintercept=0, color="blue", size=1)
)

p.show()
```

Let's say we define a category of 'big' animal as being greater than average weight and hindfoot length greater than median length and we create a new column that has an entry of 'big' if the animal meets these criteria. This example uses numpys .select() method.

In [ ]:
```
conditions = [
    (surveys['weight'] >= surveys['weight'].mean()) & (surveys['hindfoot_

choices =  ['big']
surveys["large_animal"] = np.select(conditions, choices, default="")
```

In [ ]:
```
# to display in a nice table all entries including new column
from IPython.display import display
```

```
# Display DataFrame with all columns
display(surveys)
```

In [ ]: