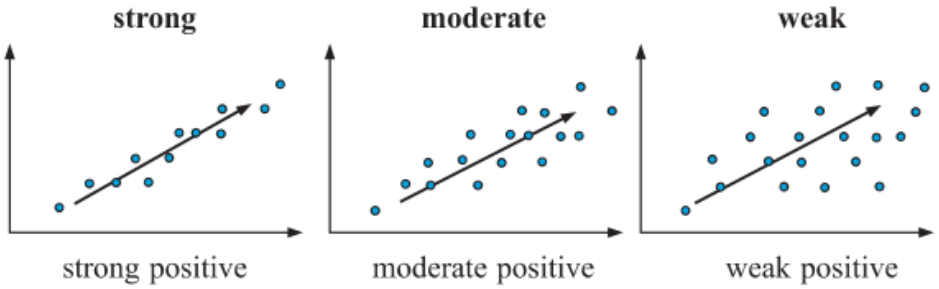
Lab 6: Correlation Analysis

1. **Generating Scatter Plots**
2. **Analyzing Correlations: Strong, Moderate, Weak, Positive, Negative**
3. **Effect of Outliers on Correlation Analysis**

Based on the textbook Statistical Thinking through Media Examples by Anthony Donoghue.

Given some dataset, we will likely want to discover any correlations (or associations) between multiple variables in the dataset. Variables may be correlated in a linear relationship, quadratic relationship, exponential relationship etc. These correlations may be positive or negative, or weak, moderate or strong. Below are three examples of variables that have a positive linear association. The strength of this linear association decreases from left to right:



It is important to note that ***correlation does not necessarily mean causation***. Two variables may be correlated, e.g. ice cream sales and sunglasses sales, but this does not mean that one causes the other, e.g. we cannot say that buying ice cream causes people to buy sunglasses.

Scatter plots allow us to visualize the relationship between variables graphically. In a scatter plot that displays the correlation between two variables, the values for one variable are plotted on the x-axis and the values for the other variable are plotted on the y-axis.

The general format for creating a scatter plot using `matplotlib` is:

```
fig, axs = plt.subplots(figsize=(10,7))
axs.scatter(x_values, y_values, c="SkyBlue")
plt.title("Plot Title", fontsize=20)
axs.set_xlabel("X-axis label", fontsize=15)
axs.set_ylabel("Y-axis label", fontsize=15)
axs.tick_params(labelsize=10)
plt.show()
```

Copy and paste this code whenever you need to create a scatter plot!

`axs.scatter(x_values, y_values, c="SkyBlue")` is the most important line of code in the block above, because it is where we give `matplotlib` the data we would like to plot. "`x_values`" and "`y_values`" are placeholders for the variables that store the list of x-values and the list of y-values respectively. The other lines of code create and format the plot.