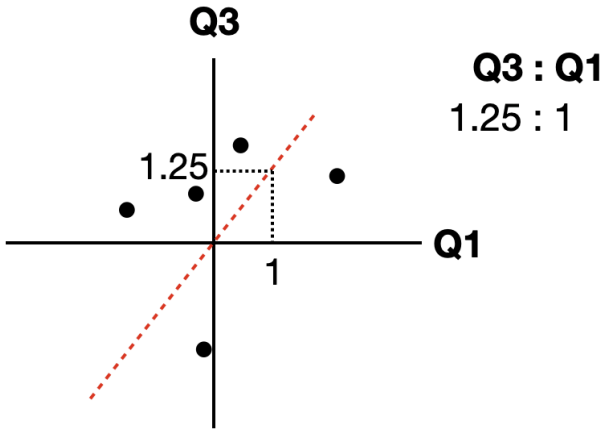
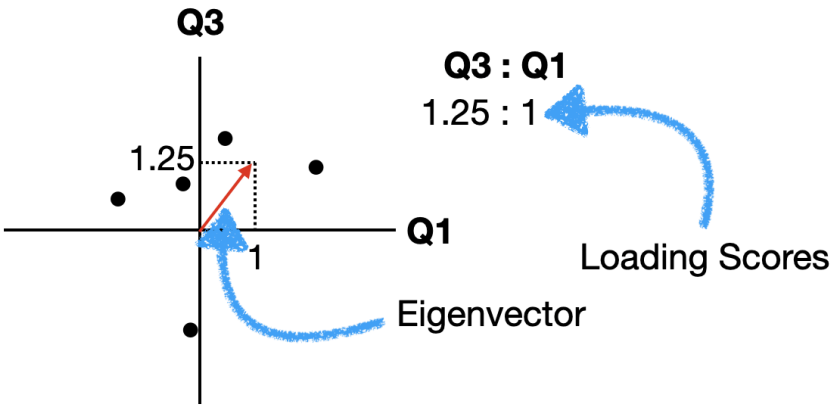
Hackathon 4: Task 3 - Principal Components

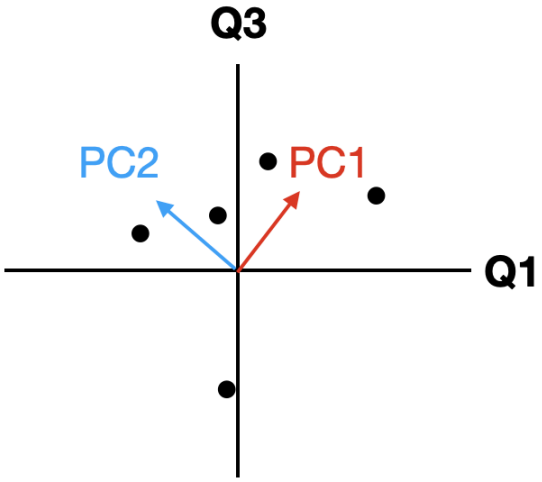
1. Choose the trend line that has the larger Eigenvalue, we will call this line the **First Principal Component** or **PC1**
2. PC1 represents the **linear combination** of your two variables, which is a fancy way of saying that it describes the mixture of the two variables, according to the ratio of rise and run (y and x). In other words, the *contribution of each variable is proportionate to its contribution to the trend line*:



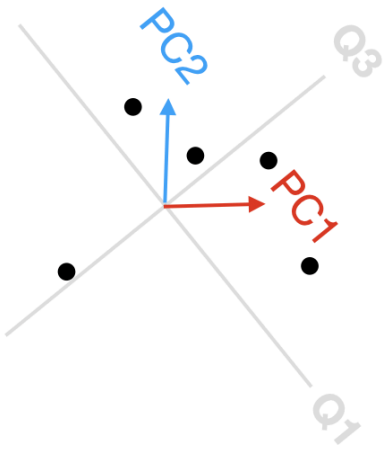
3. In Principal Component Analysis these ratios are written scaled so that they are proportionate to one unit *along the principal component*. Calculate the ratio of your two variables according to PC1 on this scale.
4. The values you have just calculated (the ratio of your two variables) are called the **Loading Scores** for PC1, the one unit long line that is comprised of the loading scores is called the **Eigenvector:**



5. Because we have two variables we can also calculate the **Second Principal Component (PC2)**. PC2 is the line perpendicular to PC1 that passes through the origin. Calculate the projected values, Eigenvalue and Eigenvector (with Loading Scores) for PC2. Note PC1 and PC2 **are not correlated** because they are perpendicular.



6. Now you have two Principal Components you can plot your data according to the Components. First, rotate the data so that PC2 is the y-axis and PC1 is the x-axis. You can do this using the projected values, it is the distance between the original and the projected values. Your plot should look like this (minus the original axes):



7. For ease of interpretation you can convert the Eigenvalues into a **Variation** metric by dividing by the sample size minus 1. Calculate variation for PC1 and PC2. The total variation is equal to variation(PC1) + variation(PC2). Therefore the relative variation that each component accounts for is variation(PC)/total variation. Plot the relative variation of each PC as a bar chart.

Check back in when your group has finished calculating the relative variation for each PC to be given task 4.