Covariance and Correlation

For the following questions, let X be an $n \times p$ real matrix.

Let R be the correlation of X. If $R = U * S * V'$, where U and V are orthogonal matrices, and S is diagonal, what is $trace(S)$?
Answer: $trace(S) = p$

Let C be the covariance matrix of X. If a and b are two scalar constants, what is the covariance of $a + bX$ in terms of C?
Answer: $cov(a + bX) = b^2C$

If C is the covariance matrix of X, and X contains one column with only zeros, how many negative eigenvalues does C have?
Answer: 0

If C is an invertible covariance matrix, is it a positive definite matrix too? If so, please provide an example; otherwise, give a counterexample
Answer: Yes. As an example, take the identity matrix I.

Let M be an $n \times p$ matrix that contains the column means of X stacked n times, and T a $p \times p$ diagonal matrix where $t_{ii} = \sigma_j$ is the standard deviation of the $j - th$ column of X. If $Y = (X - M)/T$, what is the mean and variance of the columns of Y?
Answer: $\mu_j = 0$ and $\sigma_j^2 = 1$

Covariance matrices are ellipsoids. Let C be the covariance matrix of X, whose none of its random variables (i.e. columns of X) covary. What kind of *linear transformation* does C perform on points in $p - space$? Give an example.
Answer: Since none of the X's random variables covary, C is a diagonal matrix that represents a stretching transformation. Any C being diagonal will be an example.

A car was driven on a test track for one hour at each of 5 speeds, and the gas mileage was calculated. Let X be our dataset with the following entries:

Miles/Hour	**Miles/Gallon**
20	24
30	28
40	30
50	28
60	24

What is the covariance of *Miles/Gallon* with respect to *Miles/Hour*? Do these two variables covary?

Answer: The covariance is 0 – they do not covary.

PCA

The covariance $C\begin{pmatrix} 11 & 5.5 \\ 5.5 & 4.12 \end{pmatrix}$ of the *Anscombe dataset* D_1, has the following ***eigendecomposition***

$$C = QLQ^{'} = \begin{pmatrix} -0.4849 & 0.8746 \\ 0.8746 & 0.4849 \end{pmatrix} \begin{pmatrix} 1.0775 & 0 \\ 0 & 14.0497 \end{pmatrix} \begin{pmatrix} -0.4849 & 0.8746 \\ 0.8746 & 0.4849 \end{pmatrix}^{'}$$

1. What is the axis of maximum spread in the dataset?
   Answer: Second column of Q, $\mathbf{q_2}$.

2. The **Regression Line** $y = 0.5x + 3$ does not match the axis of maximum spread of D_1. Show, mathematically, that the regression line for the *Anscombe dataset* D_1 is **NOT** parallel to its first principal component.
   Answer: The unit vector parallel to the regression line is $\mathbf{r} = \left[\frac{2}{\sqrt{5}} \quad \frac{1}{\sqrt{5}}\right]^{'}$. Then, let $\mathbf{q_2}$ be the first principal component of D_1. Next, $\langle \mathbf{r}, \mathbf{q_2} \rangle \neq \pm 1$, which proves the regression line and the principal component of D_1 are not the same.

3. What is the coordinate (i.e. *coefficient*) of the mean point in D_1, $\mu = \begin{bmatrix} 9 \\ 7.5 \end{bmatrix}$, with respect to its first principal component? You don't need to compute it numerically, but provide a mathematical expression that needs to be evaluated.
   Answer: $\langle \mu, \mathbf{q_2} \rangle = \begin{bmatrix} 9 \\ 7.5 \end{bmatrix}^{'} \begin{bmatrix} 0.8746 \\ 0.4849 \end{bmatrix}$