Exploratory Data Analysis HW: Question 1

1.1 : $m(a+bX) = a + b \times m(X)$

$m(a+bX) = \frac{1}{N}\sum_{i=1}^{N}(a+bx_i)$ Use the definition of the mean here

$m(a+bX) = \frac{1}{N}\left(\sum_{i=1}^{N} a + \sum_{i=1}^{N} bx_i\right)$

$\sum_{i=1}^{N} a = Na$ 1st part $\qquad \sum_{i=1}^{N} bx_i = b\sum_{i=1}^{N} x_i$ second part

$m(a+bX) = \frac{1}{N}\left(Na + b\sum_{i=1}^{N} x_i\right)$ combine

$m(a+bX) = a + b \cdot \frac{1}{N}\sum_{i=1}^{N} x_i$ simplify this out

$m = (a+bX) = a + b \cdot m(X)$

1.2: $cov(X, a+bY) = \frac{1}{N}\sum (x_i - m(X))[(a+by_i) - m(a+bY)]$

We already know that $m(a+bY) = a + b \cdot m(Y)$

$(a+by_i) - (a+bm(Y)) = b(y_i - m(Y))$

$cov(X, a+bY) = \frac{1}{N}\sum (x_i - m(X))\, b(y_i - m(y))$ plug it all back in

$cov(X, a+bY) = b \cdot \frac{1}{N}\sum (x_i - mX)(y_i - mY)$ pull out the "b"

$COV(X, a+bY) = b \cdot cov(X, Y)$ this is the definition!!

1.3: $var(X) = \frac{1}{N}\sum (x_i - m(X))^2$ definition of variance

$cov(X, Y) = \frac{1}{N}\sum (x_i - m(X))(y_i - m(Y))$. definition of covariance

$cov(X, X) = \frac{1}{N}\sum (x_i - m(X))(x_i - m(X))$ $y = X$

$\frac{1}{2}\sum (x_i - m(X))^2$ simplify

$Cov(X, X) = var(X)$ this is the definition!!

1.4: For any non-decreasing transformation g (like $2+5x$ or $\operatorname{arcsinh}(x)$), the median of the transformed data equals the transform of the median: $\text{median}(g(X)) = g(\text{median}(X))$; the same holds for every quantile p: $Q_{g(X)}(p) = g(Q_X(p))$ (exact when g is strictly increasing). Measures of spread generally do NOT stay the same under nonlinear g: the IQR becomes $g(Q_{0.75}) - g(Q_{0.25})$, which usually is not a simple multiple of the original IQR, and the range becomes $g(\max X) - g(\min X)$. A clean scaling happens only for linear transforms $g(x) = a+bx$ with $b > 0$: then $IQR(g(X)) = b\,IQR(X)$ and $\text{range}(g(X)) = b\,\text{range}(X)$. So, the medians/quantiles "pass through" any monotone g, while spreads only scale nicely

for straight-line transforms.

1.5: No, is not always true that $m(g(X)) = g(m(X))$. The mean of a transformed variable, $m(g(X))$, is generally not equal to applying the transform to the mean, $g(m(X))$, unless g is a straight line function (like $g(X) = a + bX$). For example, if X takes values (-1) and (1) equally, then $m(X) = 0$. If we take $g(X) = e^X$, then $g(m(X)) = e^0 = 1$, but $m(g(X)) = \frac{1}{2}(e^{-1} + e^1) \approx 1.54$, which is not equal. In general, for nonlinear functions, the average of transformed values is different from the transform of the average.