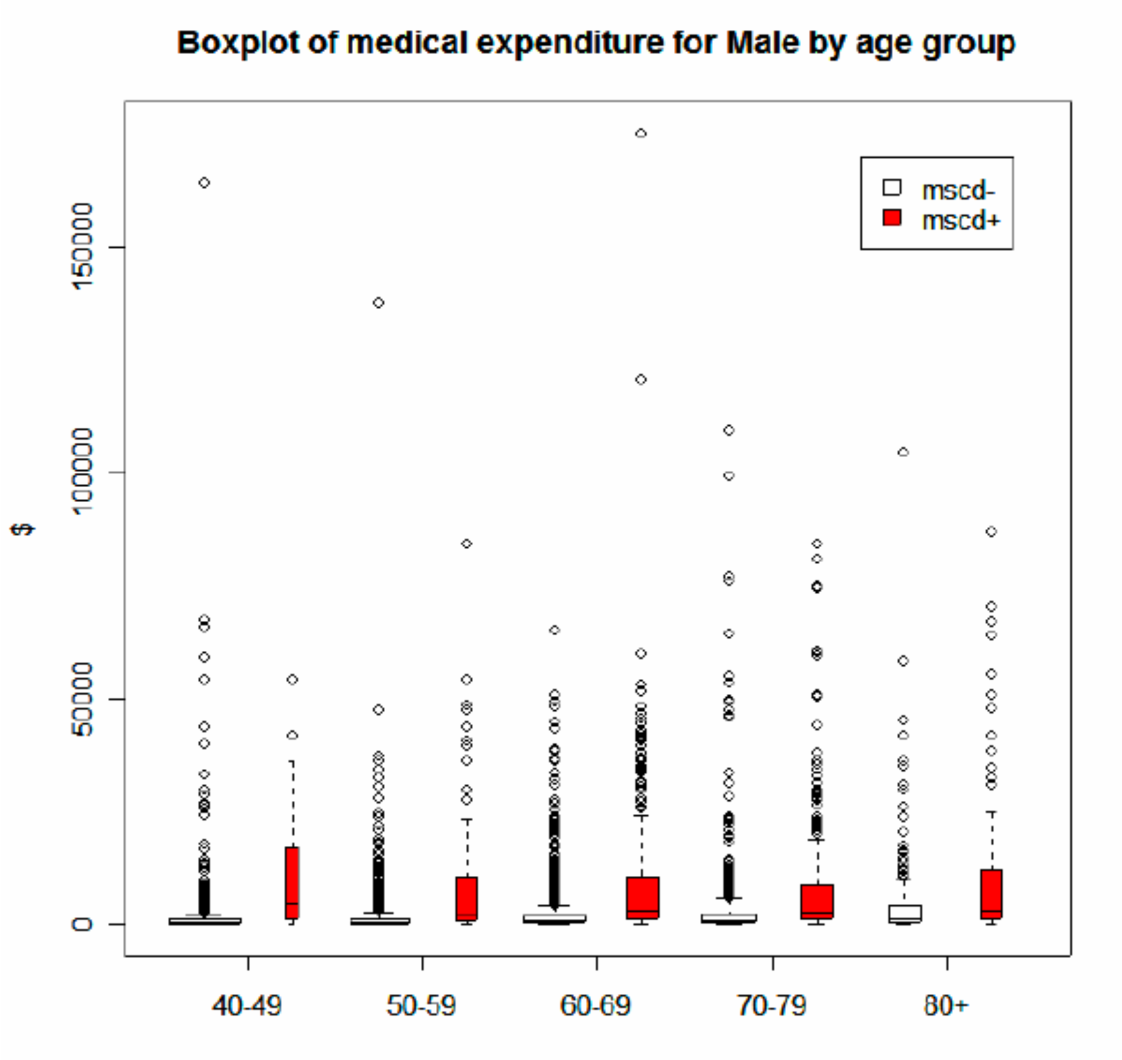
Name:

Quiz 1 (Total 40 points)

1. In R, a data analyst runs `summary(x)` on a numeric vector and obtains: Min = 40, 1st Qu. = 55, Median = 62, Mean = 65, 3rd Qu. = 74, Max = 130. (a) Compute the IQR and the upper fence for outliers. (b) Based on the fence, is the maximum value (130) an outlier? (10 points)

Solution: *(a) IQR = 74 − 55 = 19; upper fence = 74 + 1.5(19) = 74 + 28.5 = 102.5. (b) Since 130 > 102.5, the maximum is flagged as an outlier.*

2. Below is a boxplot of medical expenditure ($) by age group for males, split into those with major smoking-caused diseases (mscd+, red) and without (mscd-, white). (20 points)



Based on this figure, chose the correct statements (multiple choices):

(a) Both groups are heavily <u>left</u>-skewed.
(b) Both groups are heavily <u>right</u>-skewed.
(c) **mscd+** patients have consistently <u>higher</u> medical spending than **mscd-** patients across every single age group.

(d) **mscd+** patients have consistently lower medical spending than **mscd-** patients across every single age group.

(e) We cannot determine whether mscd+ patients have higher medical spending then mscd- patients.

(f) The overall spread of spending is elevated for people with smoking-related disease.

(g) There are more outliers (the extreme black diamonds above 100,000) appear in **mscd-** patients in 40-49 age group.

(h) There are more outliers (the extreme black diamonds above 100,000) appear in **mscd+** patients in 40-49 age group.

3. In the variance (S^2) formula below

$$S^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2,$$

explain why the term $(X_i - \overline{X})$is squared in the formula. Consider what would happen if we simply summed the deviations without squaring them. (10 points)

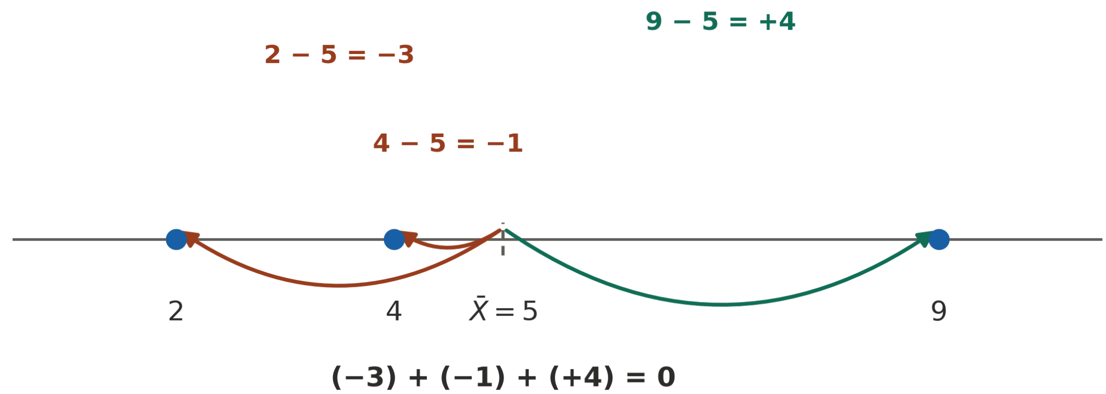
Solution:

1. Squaring prevents the deviations from canceling out.

The mean is the “balance point” of the data — the positive and negative deviations from it always sum to exactly zero, for any data set:

$$\Sigma (X_i - \bar{X}) = 0$$

Example: data = 2, 4, 9 → mean = 5. The figure below shows each deviation as an arrow from the mean to the data point:



The two negative deviations exactly cancel the one positive deviation. If we used the raw sum of deviations as our measure of spread, we would always get zero — no matter how tightly clustered or widely scattered the data are. Squaring makes every term positive before we add, so the positives and negatives can no longer cancel:

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
2	$2 - 5 = -3$	$(-3)^2 = 9$
4	$4 - 5 = -1$	$(-1)^2 = 1$
9	$9 - 5 = +4$	$(+4)^2 = 16$
Sum	0	26

2. Squaring also penalizes large deviations more heavily. (bonus)

Because squaring grows quadratically, a deviation twice as large contributes four times as much to the sum — not twice as much. A single point far from the mean pulls the variance up much more than several points only slightly off, which is exactly what a good measure of spread should do.