Assignment 1 Solutions

1. Size of car is not a quantitative variable, so it must be categorical. Since there is a natural ordering to the values of this variable, it is **categorical and ordinal**.

   Make of car is not a quantitative variable, so it must be categorical. Since there is *not* a natural ordering to the values of this variable, it is **categorical and nominal**.

   Price is a **quantitative** variable since the data is recorded as numbers and these numbers do not serve purely as labels.

2. We first find the position of the median. The median is in position $(n+1)/2 = 103/2 = 51.5$ of the ordered data values (i.e., the average of the 51^{st} and 52^{nd} ordered values).

   There are 51 data values before the median and 51 data values after the median. The third quartile is the median of the last 51 data values, so Q_3 is in position $(51+1)/2 = 26$ among the last 51 data values. In other words, Q_3 is in position $51+26 = 77$ since there are 51 data values before the median (note that the median is not one of the data values).

   There are 6 data values ≤ 25. There are $6 + 14 = 20$ data values ≤ 50. There are $6{+}14{+}17 = 37$ data values ≤ 75. There are $6{+}14{+}17{+}25 = 62$ data values ≤ 100. There are $6{+}14{+}17{+}25{+}14 = 76$ data values ≤ 125. There are $6{+}14{+}17{+}25{+}14{+}10 = 86$ data values ≤ 150.

   The 77^{th} ordered data value must therefore be contained between 125 and 150. Even though we don't know the exact value of the third quartile, we know it must lie in the $125 - 150$ interval.

   Note that, because Q_3 is in position 26 among the last 51 data values, we can also find the position of Q_3 as the 26^{th} largest data value. There is 1 data value ≥ 250, $1+1 = 2$ data values ≥ 225, $1+1+3 = 5$ data values ≥ 200, $1+1+3+5 = 10$ data values ≥ 175, and $1+1+3+5+6 = 16$ data values ≥ 150, $1+1+3+5+6+10 = 26$ data values ≥ 125. Therefore, the 26^{th} largest data value must lie in the $125 - 150$ interval.

3. We first need to find the median of this data set to help us find Q_1 and Q_3. The median is in position $(n+1)/2 = 55/2 = 27.5$, i.e., the average of the 27^{th} and 28^{th} ordered values. Therefore, the median is $(162 + 164)/2 = 163$.

   The first quartile is the median of the first 27 ordered data values, so Q_1 is in position $(27+1)/2 = 14$. Therefore, $Q_1 = 157$.

   The third quartile is the median of the last 27 ordered data values, so $Q3$ is in position $(27+1)/2 = 14$ above the median. Equivalently, we could count 14 positions down from the maximum. Therefore, $Q3 = 168$.

   Next, we calculate the lower and upper fences:

   $LF = Q_1 - 1.5IQR = Q_1 - 1.5(Q_3 - Q_1) = 157 - 1.5(168 - 157) = 157 - 16.5 = 140.5$

   $UF = Q_3 + 1.5IQR = Q_3 + 1.5(Q_3 - Q_1) = 168 + 1.5(168 - 157) = 168 + 16.5 = 184.5$

   Any data value less than 140.5 or greater than 184.5 will be labeled as an outlier. We see that there is 1 outlier on the left (137) and one outlier on the right (193).

4. The left whisker extends to the lowest data value which is not an outlier (i.e., 141). The right whisker extends to the highest data value which is not an outlier (i.e., 184).

5. Standard deviation is a measure of spread around the mean. In both data sets, the first, third, and fifth observations have the same values. In data set A, the second and fourth observations (1 and 19) are very far away from the mean. However, in data set B, the second and fourth observations (6 and 14) are not as far away from the mean.

6. John's final percentage grade is calculated as a weighted average, using the weights given to different assessment items. We are given the desired weighted average (75%), we have all of the weights needed in the calculation, and we need to solve for the score on the missing assessment item (the final exam):

$$\begin{aligned}\overline{x}_W &= \frac{x_1w_1 + x_2w_2 + x_3w_3}{w_1 + w_2 + w_3} \\ 75 &= \frac{(70)(35) + (84)(15) + (x_3)(50)}{35 + 15 + 50} \\ 7500 &= 2450 + 1260 + 50x_3 \\ x_3 &= 75.8\end{aligned}$$

7. Correlation is only defined for two quantitative variables. Since gender is a categorical variable, this is not a valid value of r.

8. The reaction time it takes to brake when driving is likely **positively** associated with amount of alcohol consumed. While it is true that amount of alcohol consumed will make you **slower** to react, slower to react means a **longer** reaction time. So, $r = -0.7$ is not a reasonable value of r.

9. People who wear larger shoes are neither more nor less likely to have a higher IQ score, so a value of $r = 0$ (indicating no association) is reasonable.

10. As distance from the equator for North American cities increases, it gets colder, in general, so average January temperature **decreases**. The sign of the correlation coefficient should be **negative**. So, $r = 0.75$ is not a reasonable value of r.

11. While there is likely to be a strong positive association between temperature and ice cream sales, a value of $r = 1.00$ implies a perfect linear relationship. A perfect linear relationship requires that **every time** temperature increases by one degree Celsius, we will see the **exact same** amount of increase in sales, not allowing for any variability in sales. A correlation of $r = 1.00$ also means that $r^2 = 1$, so, by the definition of r^2, 100% of the variation in ice cream sales would have to be explained by temperature. This is not reasonable, as there are clearly other factors besides temperature that can explain the variation in ice cream sales. For example, amount of money spent on advertising can also have an effect on ice cream sales.

12. We are given that 79.43% of the variation in gas mileage is explained by its regression on horsepower. So $r^2 = 0.7943$. The correlation between horsepower and gas mileage is

    $$r = -\sqrt{r^2} = -\sqrt{0.7943} = -0.89$$

    Note that we take the negative square root of r^2, as the relationship between horsepower and gas mileage is negative. The sign of the correlation must be the same as the sign of the slope of the least-squares regression line (the slope is -0.11).

13. The residual for Car #6 is

    $$y_6 - \hat{y}_6 = 19.2 - (35.42 - 0.11(175)) = 19.2 - 16.17 = 3.03$$

    The negative residual tells us that the point for Car #6 falls below the least-squares regression line, i.e., the gas mileage for Car #6 is lower than we would have predicted it to be, based on its horsepower.

14. For every 1 additional horsepower, the **predicted** gas mileage decreases by 0.11 miles per gallon.

15. In general, the reason there's a difference between actual y values and predicted y values (we call this difference a residual) is because the least-squares regression line is using just one x variable to help explain the variation in the y variable. We can quantify how much of the variation of the y variable is explained by its regression on the one x variable using the value of r^2.

    So, in this question, the reason there's a difference between the actual average January temperature for Winnipeg and the predicted average January temperature for Winnipeg is because we are only using latitude to explain the variation in temperature. There are other variables that are not included in the regression equation that also play a role in explaining the variation in temperature. For example, proximity to large bodies of water also affects average temperature.