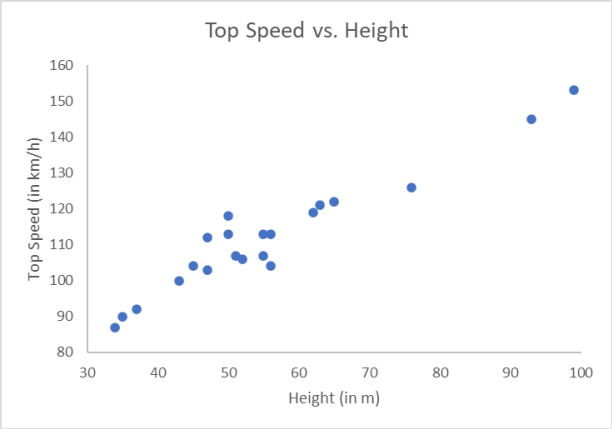
SOLUTIONS TO LAB 5 ASSIGNMENT

Question 1

As the data is obtained from observed values from 21 selected roller coasters, this is an observational study, not an experiment. No random sampling occurred since it is a list of 21 roller coasters, so population inferences are not possible. Since it is an observational study, random assignment cannot be present, so causal inferences are also not possible.

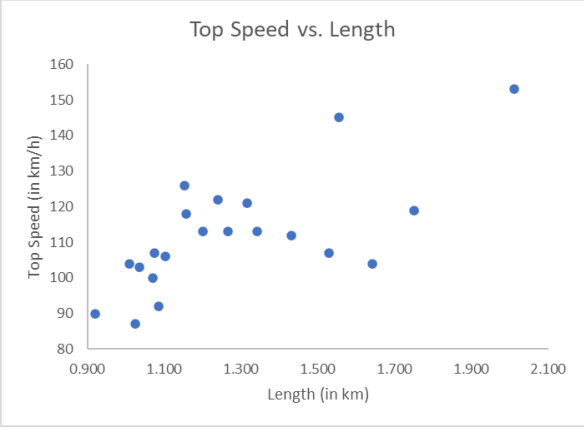
Question 2

(a) The scatterplot of *TopSpeed* (y) vs. *Height* (x) is below.



(b) The overall pattern appears to be a linear relationship. The strength of this relationship is quite strong and also has a positive association. In general, as *Height* increases, *TopSpeed* increases.

(c) The scatterplot of *TopSpeed* (y) vs. *Length* (x) is below.



The overall pattern could be seen as a linear relationship. The strength of this relationship is definitely not as strong yet still of moderate strength and also has a positive association. (Most points in the middle between 1.1530 km and 1.6410 km might suggest a negative association.) In general, as *Length* increases, *TopSpeed* increases as well.

(d) From the regression output (or using the CORREL() function), the value of the correlation coefficient between *TopSpeed* and *Height* is 0.9599 (to 4 significant decimal places). Note that Excel provides "Multiple R", which is the absolute value of the correlation. The value of the correlation coefficient between *TopSpeed* and *Length* is 0.6950 (to 4 significant decimal places).

(e) The magnitude and the sign of the first correlation coefficient are consistent with the conclusions reached above. Indeed, the linear relationship for the pair of variables is positive and the magnitude of the coefficient is as expected as it describes a strong linear relationship. The magnitude and sign are also consistent for the second correlation coefficient, though comparatively smaller in magnitude by a noticeable amount.

(f) The needed regression output is shown as part of Question 3. The percentage of variation in *TopSpeed* that is explained by its linear regression on *Height* is 92.1405%. The value is of a strong magnitude, indicating the regression model fits the data substantially well.

(g) The needed regression output is shown as part of Question 4. The percentage of variation in *TopSpeed* that is explained by its linear regression on *Length* is 48.2970%. The value is of a somewhat acceptable magnitude (though less than 50%), indicating the regression model fits the data acceptably well yet the scatterplot suggests it may not fit that well.

Question 3

The output for the regression of *TopSpeed* (*y*) on *Height* (*x*) is below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9598984
R Square	0.9214049
Adjusted R Square	0.9172683
Standard Error	4.6301356
Observations	21

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4775.2465	4775.2465	222.7452	6.0136E-12
Residual	19	407.3250	21.4382		
Total	20	5182.5714			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	60.814503	3.584512	16.96591	6.1922E-13	53.31203	68.31697
Height	0.9204914	0.061676	14.92465	6.0136E-12	0.791402	1.049581

(a) According to the output above, the equation of the least-squares regression line is as follows.

TopSpeed = 60.8145 + 0.9205**Height*

(b) Assumptions:
Linearity? The linear association between the variables is quite strong.
Normality? Aside from two outliers, normality plot appears to indicate normality. Histogram of residuals indicates slight right-skewness. Assumption mostly satisfied.
Constant variance? See below in part (d).
Random/independent observations? Random sampling did not occur, yet observations for responses appear to be independent.

H_0: $\beta_1 = 0$ vs. H_A: $\beta_1 > 0$

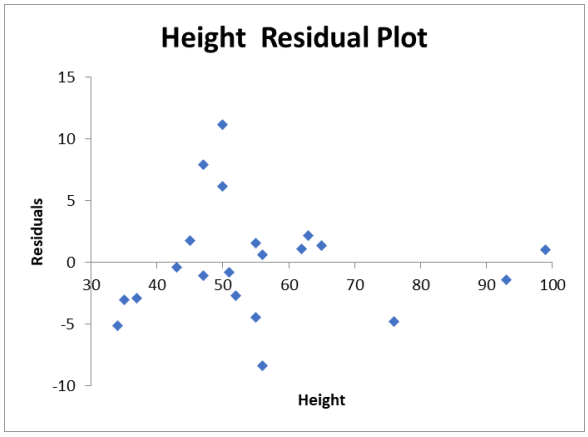
The output yields a test statistic of 14.9247 (to 4 significant decimal places). The test statistic follows a *t*-distribution with 19 degrees of freedom and a corresponding one-sided *p*-value of 3.0068×10^{-12} in scientific notation (with 4 significant decimal places). Thus, there is strong to convincing evidence against H_0 and sufficient evidence of a positive slope. (Note: not all assumptions hold so conclusions may be invalid.)

(c) The prediction can be obtained by plugging the number 50 into the equation obtained above.

TopSpeed = 60.8145 + 0.9205*(50) = 106.8391

The predicted top speed of 106.8391 km/h is somewhat close to the observed value of 113 km/h; the residual of 6.1609 available in the output reasonable in magnitude, considering the scale.

(d) The residual plot of the linear regression of *TopSpeed* (*y*) on *Height* (*x*) is below.



A residual plot is best able to identify concerns regarding the assumption of constant variance. The variance appears to decrease as height increases, indicating a funnel (or decrescendo) pattern of non-constant variance. There are some outliers in terms of leverage on the far right at (93, -1.4202) and (99, 1.0568) and their removal could potentially indicate less issues yet the non-constant variance appears it would remain. Thus, the residual plot displays moderate concerns about the constant variance assumption not being satisfied.

Question 4

The output for the regression of *TopSpeed* (*y*) on *Length* (*x*) is below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.694961
R Square	0.48297
Adjusted R Square	0.455758
Standard Error	11.87555
Observations	21

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2503.0278	2503.0278	17.74836869	0.0004713
Residual	19	2679.5436	141.0286		
Total	20	5182.5714			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	61.04795	12.40203	4.922417	9.45617E-05	35.09021	87.00569
Length	39.86302	9.462176	4.212881	0.000471263	20.05846	59.66759

(a) According to the output above, the equation of the least-squares regression line is as follows.

TopSpeed = 61.0480 + 39.8630**Length*

(b) Assumptions:
Linearity? The linear association between the variables is moderately strong with noted variation.
Normality? Aside from two outliers, normality plot appears to indicate normality. Histogram of residuals indicates left-skewness. Assumption reasonably satisfied.
Constant variance? See below in part (d).
Random/independent observations? Random sampling did not occur, yet observations for responses appear to be independent.

H_0: $\beta_1 = 0$ vs. H_A: $\beta_1 > 0$

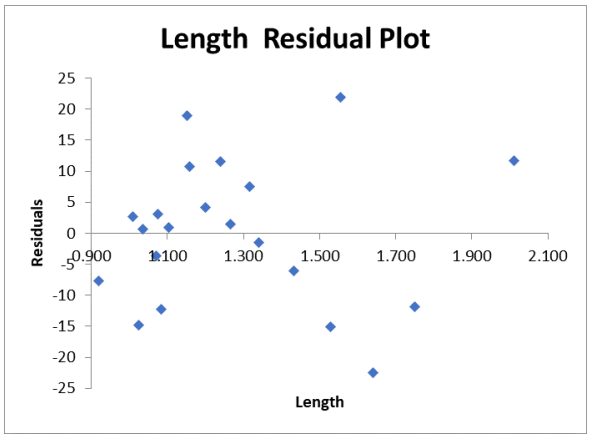
The output yields a test statistic of 4.2129 (to 4 significant decimal places). The test statistic follows a *t*-distribution with 19 degrees of freedom and a corresponding one-sided *p*-value of 0.0002356 (with 4 significant decimal places). Thus, there is strong to convincing evidence against H_0 and sufficient evidence of a positive slope. (Note: not all assumptions hold so conclusions may be invalid.)

(c) The prediction can be obtained by plugging the number 1.200 into the equation obtained above.

TopSpeed = 61.0480 + 39.8630*(1.200) = 108.8836

The predicted top speed of 108.8836 km/h is quite close to the observed value of 113 km/h; the residual of 4.1164 available in the output is quite small in magnitude, considering the scale. Incidentally, this is a lower residual than the one in Question 3.

(d) The residual plot of the linear regression of *TopSpeed* (y) on *Length* (x) is below.



A residual plot is best able to identify concerns regarding the assumption of constant variance. There are some relative outliers in the top right at (2.012, 11.7476) and (1.555, 21.9650), yet they are arguably assisting the constant variance assumption as their absence could indicate less constant variance across all values of length. Nevertheless, the clear pattern presents slight to moderate concerns about the constant variance assumption not being satisfied.

Question 5

When comparing the two models, the regression model between *Height* and *TopSpeed* has stronger linearity, a larger R^2 value (and correlation by extension), though not better overall satisfaction of assumptions. Still, based off this data, the first regression model is better. If an engineer wanted to design a new roller coaster to maximize top speed, height appears more reliable to measure top speed rather than length. It is worth noting that no population inferences are possible for this data, so this conclusion is confined to this dataset.

LAB 5 ASSIGNMENT MARKING SCHEMA

Question 1 (6)

Type of study: 2 points
Population inferences: 2 points
Causal inferences: 2 points

Question 2 (32)

(a) Scatterplot of x_1 vs. y: 6 points
(b) Description of the pattern: 3 points
(c) Scatterplot of x_2 vs. y: 6 points
Description of the pattern: 3 points
(d) Correlation: 2 points each (4 points total)
(e) Comparison with scatterplots: 2 points each (4 points total)
(f) Percentage of variation: 2 points
Regression model fit: 1 point
(g) Percentage of variation: 2 points
Regression model fit: 1 point

Question 3 (30)

(a) Regression output: 6 points
Equation of least-squares regression: 3 points
(b) Test for positive slope: 8 points
(assumptions: 2, hypotheses: 1, distribution: 1, test statistic: 2, *p*-value: 1, conclusion: 1)
(c) Prediction: 2 points
Comparison: 2 points
(d) Residual plot: 6 points
Description of the pattern and relation to assumptions: 3 points

Question 4 (30)

(a) Regression output: 6 points
Equation of least-squares regression: 3 points
(b) Test for positive slope: 8 points
(assumptions: 2, hypotheses: 1, distribution: 1, test statistic: 2, *p*-value: 1, conclusion: 1)
(c) Prediction: 2 points
Comparison: 2 points
(d) Residual plot: 6 points
Description of the pattern and relation to assumptions: 3 points

Question 5 (6)

Summary: 2 points
Best model: 2 points
Explanatory variable of focus: 2 points

TOTAL = 104