Part II	**Exploring Relationships Between Variables**
Chapter 7	**Scatterplots, Association, and Correlation**
Scatterplot _____ is plotted on the x-axis. _____ is plotted on the y-axis.	Shows the relationship between two quantitative variables on the same cases (individuals). Explanatory (*independent*/input) variable Response (*dependent*/output) variable
Once we make a scatterplot, we describe association by telling about:	**1. Form**: straight, curved, no pattern, other? **2. Direction**: + or – slope? **3. Strength**: how much scatter {how closely points follow the form} **4. Unusual Features**: outliers, clusters, subgroups?
_____ is a deliberately vague term describing the relationship between two variables. If positive then _____	Association increases in one variable generally correspond to increases in the other.
Correlation describes the ____ and ____ of the ____ relationship between two _____ variables, without significant _____	strength direction, linear quantitative outliers.
3 conditions needed for Correlation:	1. Quantitative Variables 2. Straight Enough 3. Outlier
The correlation coefficient is found by _____. It's value ranges from _____ , it has no _____, and is immune to changes of _____	finding the average product of the z-scores (standardized values). $r = \frac{\sum z_x z_y}{n-1}$ -1 to +1 units. scale or order.
Perfect correlation r = _____, occurs only when _____.	± 1 the points lie exactly on a straight line. (you can perfectly predict one variable knowing the other)
No correlation r = _____, means that knowing one variable gives you _____	0 no information about the other variable.
You should give the _____ and _____ of x and y along with the correlation because …	Mean Standard deviation Correlation is not a complete description of two-variable data and the its formula uses means and standard deviations in the z-scores.
Scatterplots and correlation coefficients never prove _____	 causation.
Lurking variable	A variable other than x and y that simultaneously affects both variables, accounting for the correlation between the two.
To add a categorical variable to an existing scatterplot _____	 use a different plot color or symbol for each category.

Chapter 8	**Linear Regression**
Regression to the mean	Because the correlation is always less than 1.0 in magnitude, each predicted $\hat{y}$ tends to be fewer standard deviations from its mean than its corresponding x was from its mean. ($\hat{z}_y = rz_x$)
Residual If positive If negative	Observed value – predicted value $y - \hat{y}$ Then the model makes an underestimate. Then the model makes an overestimate.
Regression line Line of best fit For standardized values For actual x and y values	The unique line that minimizes the variance of the residuals (sum of the squared residuals). $\hat{z}_y = rz_x$ $\hat{y} = b_0 + b_1x$
To calculate the regression line in real units (actual x and y values)	1. Find slope, $b_1 = \frac{rs_y}{s_x}$ 2. Find y-intercept, plug b_1 and point (x, y) [usually $(\bar{x}, \bar{y})$] into $\hat{y} = b_0 + b_1x$ and solve for b_0 3. Plug in slope, b_1, and y-intercept, b_0, into $\hat{y} = b_0 + b_1x$
3 conditions needed for Linear Regression Models: /* same as correlation */	1. Quantitative Variables 2. Straight Enough – check original scatterplot & residual scatterplot 3. Outlier (clusters) –points with large residuals and/or high leverage
R^2	The square of the correlation, r, between x and y The success of the regression model in terms of the fraction of the variation of y accounted for by the model. (differences in x explain XX% of the variability in y) (The model explains XX% of the variability in y)
A high R^2	Does not demonstrate the appropriateness of the regression.
Looking at a _____ is a good way to check the Straight Enough Condition. It should be _____	a scatterplot of the residuals vs. the x-values. (appropriateness) boring: uniform scatter with no direction, shape, or outliers..
The ____ is the key to assessing how well the model fits (extracts the form).	variation in the residuals
Standard deviation of the residuals, s_e	Gives a measure of how much the points spread around the regression line.
$1 - R^2$	The fraction of the original variation left in the residuals. (The percentage of variability not explained by the regression line.)
Extrapolations	Dubious predictions of y-values based on x-values outside the range of the original data.