APPENDIX A

DATA SET DESCRIPTIONS

Anscombe

This data set is used to illustrate the importance of statistical display as an adjunct to summary statistics. Anscombe (1973) fabricated four different bivariate data sets such that, for all data sets, the respective X and Y means, X and Y standard deviations, and correlations, slopes, intercepts, and standard errors of estimate are equal. Accordingly, without a visual representation of these four panels, one might assume that the data values for all four data sets are the same. Scatterplots illustrate, however, the extent to which these data sets are different from one another.

Basketball

The data set consists of the heights and weights of the 24 scoring leaders, 12 each from the U.S. Women's and Men's National Basketball Association, for the 2014 – 2015 season. These data are taken from the ESPN website at espn.go.com.

Variable Name	Variable Label	Additional Description
PLAYER	Player name	0 = Male; 1 = Female
GENDER		
HEIGHTIN	Height in inches	
WEIGHTLB	Weight in pounds	
GAMES	Games played	
MINUTESGAME	Minutes per game	
POINTS	Average total points scored per game	

Blood

The data were collected to determine whether an increase in calcium intake reduces blood pressure among African-American adult males. The data are based on a sample of 21 African-American adult males selected randomly from the population of African-American adult males. These data come from the Data and Story Library (DASL) website. Ten of the 21 men were randomly assigned to a treatment condition that required them to take a calcium supplement for 12 weeks. The remaining 11 men received a placebo for the 12 weeks. At both the beginning and the end of this time period, systolic blood pressure readings of all men were recorded.

Variable Name	Variable Label	Additional Description
ID		
SYSTOLC1	Initial blood pressure	
SYSTOLC2	Final blood	

	pressure	
TREATMEN	Treatment	0 = Placebo; 1 = Calcium

Brainsz

The data set and description are taken from the DASL (Data Sets and Story Library) website. The data are based on a study by Willerman et al. (1991) of the relationships between brain size, gender, and intelligence. The research participants consisted of 40 right-handed introductory psychology students with no history of alcoholism, unconsciousness, brain damage, epilepsy, or heart disease who were selected from a larger pool of introductory psychology students with total Scholastic Aptitude Test Scores higher than 1350 or lower than 940. The students in the study took four subtests (Vocabulary, Similarities, Block Design, and Picture Completion) of the Wechsler (1981) Adult Intelligence Scale-Revised. Among the students with Wechsler full-scale IQ's less than 103, 10 males and 10 females were randomly selected. Similarly, among the students with Wechsler full-scale IQ's greater than 130, 10 males and 10 females were randomly selected, yielding a randomized blocks design. MRI Scans were performed at the same facility for all 40 research participants to measure brain size. The scans consisted for 18 horizontal MRI images. The computer counted all pixels with non-zero gray scale in each of the 18 images and the total count served as an index for brain size.

Variable Name	Variable Label	Additional Description
ID		
GENDER		0 = Male; 1 = Female
FSIQ	Full Scale IQ Score based on WAIS-R	
VIQ	Verbal IQ Score	
PIQ	Performance IQ Score	
MRI		
IQDI		0 = Lower IQ; 1 = Higher IQ

Currency

This data set contains, for the smaller bill denominations, the value of the bill and the total value in circulation. The source for these data is *The World Almanac and Book of Facts 2014*.

Variable Name	Variable Label
BILLVALU	Denomination
CIRC	Total currency in circulation

Exercise, Food Intake and Weight Loss

A fabricated data set constructed by Darlington (1990) to demonstrate the importance of including all relevant variables in an analysis. This data set contains information about exercise, food intake, and weight loss for a fictional set of dieters.

Variable Name	Variable Label	Additional Description
FOOD	Food intake	The average daily number of calories consumed in one particular week that is more than a baseline of 1,000 calories as measured in increments of 100 calories
WEIGHTLOSS	Weight loss	The number of pounds lost in that week
EXERCISE	Exercise	The average daily number of hours exercised in that week

Framingham

The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of non-institutionalized people in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. The study began in 1956 and 5,209 subjects were initially enrolled in the study. In our data set, we included variables from the first examination in 1956 and the third examination, in 1968. Clinic examination data has included cardiovascular disease risk factors and markers of disease such as blood pressure, blood chemistry, lung function, smoking history, health behaviors, ECG tracings, Echocardiography, and medication use. Through regular surveillance of area hospitals, participant contact, and death certificates, the Framingham Heart Study reviews and adjudicates events for the occurrence of any of the following types of coronary heart disease(CHD): Angina Pectoris, Myocardial Infarction, Heart Failure, and Cerebrovascular disease.

The associated dataset is a subset of the data collected as part of the Framingham study and includes laboratory, clinic, questionnaire, and adjudicated event data on 400 participants. These participants for the data set have been chosen so that among all male participants, 100 smokers and 100 non-smokers were selected at random. A similar procedure resulted in 100 female smokers and 100 female non-smokers. This procedure resulted in an over-sampling of smokers. The data for each participant is on one row. People who had any type of CHD in the initial examination period are not included in the dataset.

Variable Name	Variable Label	Additional Description
ID		
SEX	Sex	1 = Men; 2 = Women
TOTCHOL1	Serum Cholesterol mg/dL (1)	

AGE1	Age (years) at examination (1)	
SYSBP1	Systolic BP mmHg (1)	
DIABP1	Diastolic BP mmHg (1)	
CURSMOKE1	Current Cig Smoker Y/N (1)	0 = No; 1 = Yes
CIGPDAY1	Cigarettes per day (1)	
BMI1	Body Mass Index (kg/(M*M) (1)	
DIABETES1	Diabetic Y/N (1)	0 = Not a diabetic; 1 = Diabetic
BPMEDS1	Anti-hypertensive meds Y/N (1)	0 = Not currently used; 1 = Currently used
HEARTRTE1	Ventricular Rate (beats/min) (1)	
GLUCOSE1	Casual Glucose mg/dL (1)	
PREVCHD1	Prevalent CHD (MI,AP,CI) (1)	0 = Free of CHD; 1 = Prevalence of CHD
TIME1	Days since Index Exam (1)	
TIMECHD1	Days Baseline-Inc Any CHD (1)	
TOTCHOL3	Serum Cholesterol mg/dL (3)	
AGE3	Age (years) at examination (3)	
SYSBP3	Systolic BP mmHg (3)	
DIABP3	Diastolic BP mmHg (3)	
CURSMOKE3	Current Cig Smoker Y/N (3)	0 = No; 1 = Yes
CIGPDAY3	Cigarettes per day (3)	
BMI3	Body Mass Index (kg/(M*M) (3)	
DIABETES3	Diabetic Y/N (3)	0 = Not a diabetic; 1 = Diabetic
BPMEDS3	Anti-hypertensive meds Y/N (3)	0 = Not currently used; 1 = Currently used
HEARTRTE3	Ventricular Rate (beats/min) (3)	
GLUCOSE3	Casual Glucose mg/dL (3)	
PREVCHD3	Prevalent CHD (MI,AP,CI) (3)	0 = Free of CHD; 1 = Prevalence of CHD
TIME3	Days since Index Exam (3)	
HDLC3	HDL Cholesterol mg/dL (3)	
LDLC3	LDL Cholesterol mg/dL (3)	
TIMECHD3	Days Baseline-Inc Any CHD (3)	
ANYCHD4	Incident Hosp MI, AP, CI, Fatal CHD by the end of the study	0 = CHD event did not occur; 1 = CHD even did occur

Hamburg

This data set contains the fat grams and calories associated with the different types of hamburger sold by McDonald's. The data are from McDonald's Nutrition Information Center.

Variable Name	Variable Label	Additional Description
CALORIES		
CHEESE	Cheese Added?	0 = No; 1 = Yes
FAT	Grams of fat	
NAME	Type of burger	

Ice Cream

This data set contains fabricated data for the temperature, relative humidity and ice cream sales for 30 days randomly selected between May 15th and September 6th.

Variable Name	Variable Label	Additional Description
ID		
BARSOLD	Number of ice cream bars sold	
TEMP	Temperature in degrees Fahrenheit	
RELHUMID	Relative humidity	

Impeach

On February 12, 1999, for only the second time in the nation's history, the U.S. Senate voted on whether to remove a president, based on impeachment articles passed by the U.S. House. Professor Alan Reifman of Texas Tech University created the data set consisting of descriptions of each senator that can be used to understand some of the reasons that the senators voted the way they did. The data are taken from the Journal of Statistics Education [online].

Variable Name	Variable Label	Additional Description
NAME	Senator's name	
STATE	State the senator represents	
REGION		1 = Northeast; 2 = Midwest; 3 = South; 4 = West
VOTE1	Vote on perjury	0 = Not guilty; 1 = Guilty
VOTE2	Vote on obstruction of justice	0 = Not guilty; 1 = Guilty
GUILTY	Number of guilty votes	0 = Democrat; 1 = Republican
PARTY		
CONSERVA	Conservatism score	The senator's degree of ideological conservatism is based on 1997 voting records as judged by the American Conservative Union, where the scores

		ranged from 0 to 100 and 100 is most conservative
SUPPORTC	State voter support for Clinton	The percent of the vote Clinton received in the 1996 Presidential election in the senator's state
REELECT	Year the senator's seat is up for reelection	2000, 2002, or 2004
NEWBIE	First-term senator?	0 = No; 1 = Yes

Learndis

What is the profile of students classified as learning disabled? The questions on this exam relate to a subset of data from a study by Susan Tomasi and Sharon L. Weinberg (1999). According to Public Law 94.142, enacted in 1976, a team may determine that a child has a learning disability (LD) if a severe discrepancy exists between a child's actual achievement in, for example, math or reading, and his or her intellectual ability. The data set consists of six variables, described below, on 105 elementary school children from an urban area who were classified as LD and who, as a result, had been receiving special education services for at least three years. Of the 105 children, 42 are female and 63 are male. There are two main types of placements for these students: part-time resource room placements in which the students get additional instruction to supplement regular classroom instruction and self-contained classroom placements in which students are segregated full time. In this data set 66 students are in resource room placements while 39 are in self-contained classroom placements. For inferential purposes, we consider the children in the data set to be a random sample of all children attending public elementary school in a certain city who have been diagnosed with learning disabilities. Many students in the data set have missing values for either math or reading comprehension, or both. Such omissions can lead to problems when generalizing results. There are statistical remedies for missing data that are beyond the scope of this text. In this case, we will assume that there is no pattern to the missing values, so that our sample is representative of the population.

Variable Name	**Variable Label**	**Additional Description**
GENDER	Gender	0 = Male; 1 = Female
GRADE	Grade level	1, 2, 3, 4, or 5
IQ	Intellectual ability	Higher scores ↔ greater IQ Scores could range from 0 to 200
MATHCOMP	Math comprehension	Higher scores ↔ greater math comprehension Scores could range from 0 to 200
PLACEMEN	Type of placement	0 = Part-time in resource room; 1 = Full-time in self-contained classroom
READCOMP		Higher scores ↔ greater reading comprehension

		Scores could range from 0 to 200

Mandex

This fictional data set contains the treatment group number and the manual dexterity scores for 30 individuals selected by the director of a drug rehabilitation center. There are three treatments and the individuals are randomly assigned ten to a treatment. After five weeks of treatment, a manual dexterity test is administered for which a higher score indicates greater manual dexterity.

Variable Name	**Variable Label**
SCORE	
TREATMEN	Manual Dexterity Score

Marijuana

The data set contains the year and percentage of twelfth graders who have ever used marijuana for several recent years. The source for these data is *The World Almanac and Book of Facts 2014*.

Variable Name	**Variable Label**
YEAR	
MARIJ	Percentage of twelfth graders who reported that they have ever used marijuana

NELS

In response to pressure from federal and state agencies to monitor school effectiveness in the United States, the National Center of Education Statistics (NCES) of the U.S. Department of Education conducted a survey in the spring of 1988. The participants consisted of a nationally representative sample of approximately 25,000 eighth graders to measure achievement outcomes in four core subject areas (English, history, mathematics, and science), in addition to personal, familial, social, institutional, and cultural factors that might relate to these outcomes. Details on the design and initial analysis of this survey may be referenced in Horn, Hafner, and Owings (1992). A follow-up of these students was conducted during tenth grade in the spring of 1990; a second follow-up was conducted during the twelfth grade in the spring of 1992; and, finally, a third follow-up was conducted in the spring of 1994.

For this book, we have selected a sub-sample of 500 cases and 50 variables. The cases were sampled randomly from the approximately 5,000 students who responded to all four administrations of the survey, who were always at grade level (neither repeated nor skipped a grade) and who pursued some form of post-secondary education. The particular

variables were selected to explore the relationships between student and home background variables, self-concept, educational and income aspirations, academic motivation, risk-taking behavior, and academic achievement.

Variable Name	Variable Label	Additional Description
ABSENT12	Number of Times Missed School	0 = Never; 1 = 1-2 Times; 2 = 3-6 Times; 3 = 7-9 Times; 4 = 10-15 Times; 5 = Over 15 Times
ACHMAT08	Math Achievement in Eighth Grade	Similar to ACHRDG08
ACHMAT10	Math Achievement in Tenth Grade	Similar to ACHRDG08
ACHMAT12	Math Achievement in Twelfth Grade	Similar to ACHRDG08
ACHRDG08	Reading Achievement in Eighth Grade	Gives a score for the student's performance in eighth grade on a standardized test of reading achievement. Actual values range from 36.61 to 77.2, from low to high achievement. 99.98 = Missing; 99.99 = Missing
ACHRDG10	Reading Achievement in Tenth Grade	Similar to ACHRDG08
ACHRDG12	Reading Achievement in Twelfth Grade	Similar to ACHRDG08
ACHSCI08	Science Achievement in Eighth Grade	Similar to ACHRDG08
ACHSCI10	Science Achievement in Tenth Grade	Similar to ACHRDG08
ACHSCI12	Science Achievement in Twelfth Grade	Similar to ACHRDG08
ACHSLS08	Social Studies Achievement in Eighth Grade	Similar to ACHRDG08
ACHSLS10	Social Studies Achievement in Tenth Grade	Similar to ACHRDG08
ACHSLS12	Social Studies Achievement in Twelfth Grade	Similar to ACHRDG08
ADVMATH8	Advanced Math Taken in Eighth Grade	0 = No; 1 = Yes; 8 = Missing
ALCBINGE	Binged on Alcohol Ever?	0 = Never; 1 = Yes
ALGEBRA8	Algebra Taken in Eighth Grade?	0 = No; 1 = Yes
APOFFER	Number of Advanced	98 = Missing

	Placement Courses Offered by School	
APPROG	Advanced Placement Program Taken?	0 = No; 1 = Yes; 6 = Missing; 8 = Missing
CIGARETT	Smoked Cigarettes Ever?	0 = Never; 1 = Yes
COMPUTER	Computer Owned by Family in 8th Grade?	0 = No; 1 = Yes
CUTS12	Number of Times Skipped/Cut Classes in Twelfth Grade	0 = Never; 1 = 1-2 Times; 2 = 3-6 Times; 3 = 7-9 Times; 4 = 10-15 Times; 5 = Over 15 Times
EDEXPECT	Highest level of education expected	1 = Less than College Degree; 2 = Bachelor's Degree; 3 = Master's Degree; 4 = Ph.D., MD, JD, etc.
EXCURR12	Time Spent Weekly on Extra-Curricular Activities in Twelfth Grade	0 = None; 1 = Less than 1 Hour; 2 = 1 - 4 Hours; 3 = 5 - 9 Hours; 4 = 10 - 14 Hours; 5 = 15 - 19 Hours; 6 = 20 - 24 Hours; 7 = 25 or More Hours
EXPINC30	Expected income at age 30	-6 = Missing, -9 = Missing
FAMSIZE	Family Size	98 = Missing
FINANAID	Received Financial Aid in College	0 = No; 1 = Yes; -9 = Missing; -8 = Missing; -7 = Missing; -6 = Missing
GENDER	Gender	0 = Male; 1 = Female
HOMELANG	Home Language Background	1 = Non-English Only; 2= Non-English Dominant; 3 = English Dominant; 4 = English Only
HSPROG	Type of High School Program	1= Rigorous Academic; 2 = Academic; 3 = Some Vocational; 4 = Other
HWKIN12	Time Spent on Homework Weekly in School Per Week in Twelfth Grade	0 = None; 1 = Less than 1 Hour; 2 = 1-3 Hours; 3 = 4-6 Hours; 4 = 7-9 Hours; 5 = 10-12 Hours; 6 = 13-15 Hours; 7 = 16-20 Hours; 8 = Over 20 Hours; 98 = Missing
HWKOUT12	Time Spent on Homework out of School per Week in Twelfth Grade	0 = None; 1 = Less than 1 Hour; 2 = 1-3 Hours; 3 = 4-6 Hours; 4 = 7-9 Hours; 5 = 10-12 Hours; 6 = 13-15 Hours; 7 = 16-20 Hours; 8 = Over 20 Hours; 98 = Missing
ID	Case Number	
IMPTEDUC	How Important is a Good Education?	1 = Not Important; 2 = Somewhat Important; 3 = Very Important
LATE12	Number Times Late for School in Twelfth Grade	0 = Never; 1 = 1-2 Times; 2 = 3-6 Times; 3 = 7-9 Times; 4 = 10-15 Times; 5 = Over 15 Times

MARIJUAN	Smoked Marijuana Ever?	0 = Never; 1 = Yes
NUMINST	Number post-secondary institutions attended	
NURSERY	Nursery School Attended?	0 = No; 1 = Yes; 8 = Missing
PARMARL8	Parents' Marital Status in Eighth Grade	1 = Divorced; 2 = Widowed; 3 = Separated; 4= Never Married; 5 = Marriage-Like Relationship; 6 = Married; 98 = Missing
REGION	Geographic Region of School	1= Northeast; 2 = North Central; 3 = South; 4 = West
SCHATTRT	School Average Daily Attendance Rate	Gives the average daily attendance rate for the secondary school the student attends. 998 = Missing
SCHTYP8	School Type in Eighth Grade	Classifies the type of school each student attended in eighth grade where 1 = Public; 2 = Private, Religious; 3 = Private, Non-Religious
SES	Socio-Economic Status	Gives a score representing the socio-economic status (SES) of the student, a composite of father's education level, mother's education level, father's occupation, mother's education, and family income. Values range from 0 to 35, from low to high SES.
SLFCNC08	Self-Concept in Eighth Grade	Similar to SLFCNC12
SLFCNC10	Self Concept in Tenth Grade	Similar to SLFCNC12
SLFCNC12	Self Concept in Twelfth Grade	Gives a score for student self-concept in twelfth grade. Values range from 0 to 43. Self-concept is defined as a person's self-perceptions, or how a person feels about himself or herself. Four items comprise the self-concept scale in the NELS questionnaire: I feel good about myself; I feel I am a person of worth, the equal of other people; I am able to do things as well as most other people; and, on the whole, I am satisfied with myself. A self-concept score, based on the sum of scores on each of these items, is used to measure self-concept in the NELS study. Higher scores associate with higher

		self-concept and lower scores associate with lower self-concept.
TCHERINT	My Teachers are Interested in Students	Classifies student agreement with the statement "My teachers are interested in students" using the Likert scale 1 = Strongly agree; 2 = Agree; 3 = disagree; 4 = strongly disagree
UNITCALC	Units in Calculus (NAEP)	Number of years of Calculus taken in high school
UNITENGL	Units in English (NAEP)	Number of years of English taken in high school
UNITMATH	Units in Mathematics (NAEP)	Number of years of Math taken in high school
URBAN	Urbanicity	Classifies the type of environment in which each student lives where 1 = "Urban", 2 = "Suburban", and 3 = "Rural".
ABSENT12	Number of Times Missed School	

Skulls

Four size measurements are taken on 150 male skulls, 30 from each of 5 different time periods between 4000 b.c. and 150 a.d. The data include the time period and the four size measurements. These data come from the Data and Story Library (DASL) website.

Variable Name	Variable Label	Additional Description
BH	Basibregmatic Height of Skull	
BL	Basialveolar Length of Skull	
MB	Maximal Breadth of Skull	
NH	Nasal Height of Skull	
YEAR	Approximate Year of Skull Formation	

States

This data set includes different measures of the 50 states and Washington, D.C. These data are from *The 2014 World Almanac and Book of Facts*.

Variable Name	Variable Label	Additional Description

EDUCEXPE	Expenditure per Pupil, on Average, 2011 - 2012	
ENROLLMT	Total Public School Enrollment 2011 - 2012	
PERTAK	Percentage of Eligible Students Taking the SAT 2012	
REGION	The region of the country in which the state is located	1 = Northeast; 2 = Midwest; 3 = South; 4 = West
SATCR	Average SAT Critical Reading 2013	
SATM	Average SAT Math 2013	
SATW	Average SAT Writing 2013	
STATE	Name of state	
STUTEACH	Pupils per Teacher, on Average 2011 - 2012	
TEACHPAY	Average Annual Salary for Public School Teachers 2011 - 2012	

Stepping

The data set and description are taken from the DASL (Data Sets and Story Library) website. Students at Ohio State University conducted an experiment in the fall of 1993 to explore the nature of the relationship between a person's heart rate and the frequency at which that person stepped up and down on steps of various heights. The response variable, heart rate, was measured in beats per minute. For each person, the resting heart rate was measured before a trial (RestHR) and after stepping (HR). There were two different step heights (HEIGHT): 5.75 inches (coded as 0), and 11.5 inches (coded as 1). There were three rates of stepping (FREQUENCY): 14 steps/min. (coded as 0), 21 steps/min. (coded as 1), and 28 steps/min. (coded as 2). This resulted in six possible height/frequency combinations. Each subject performed the activity for three minutes. Subjects were kept on pace by the beat of an electric metronome. One experimenter counted the subject's heart rate, in beats per minute, for 20 seconds before and after each trial. The subject always rested between trials until her or his heart rate returned to close to the beginning rate. Another experimenter kept track of the time spent stepping. Each subject was always measured and timed by the same pair of experimenters to reduce variability in the experiment.

Variable Name	**Variable Label**	**Additional Description**
ORDER	The overall performance order of the trial	
BLOCK	The subject and experimenters' block number	
HEIGHT	Step Height	0 = Low; 1 = High
FREQ	Rate of stepping	0 = Slow; 1 = Medium; 2 = Fast
HRINIT	The resting heart rate of the subject before a trial,	

	in beats per minute	
HRFINAL	The final heart rate of the subject after a trial, in beats per minute	

Temp

This data set gives the average monthly temperatures (in Fahrenheit) for Springfield, MO and San Francisco, CA. These data are from Burrill and Hopensperger (1993).

Variable Name	Variable Label	Additional Description
CITY		1 = Springfield, MO; 2 = San Francisco, CA
TEMP	Average monthly temperature	

Wages

This is a subsample of 100 males and 100 females randomly selected from the 534 cases that comprised the 1985 Current Population Survey in a way that controls for highest education level attained. The sample of 200 contains 20 males and 20 females with less than a high school diploma, 20 males and 20 females with a high school diploma, 20 males and 20 females with some college training, 20 males and 20 females with a college diploma, and 20 males and 20 females with some graduate school training. The data include information about gender, highest education level attained, and hourly wage.

Variable Name	Variable Label	Additional Description
ED	Highest education level	1 = Less than a HS degree; 2 = HS degree; 3 = Some college; 4 = College degree; 5 = Graduate school
EDUC	Number of years of education	
EXPER	Number of years of work experience	
ID		
MARR	Marital status	0 = Not married; 1 = Married
OCCUP		1 = Management; 2 = Sales; 3 = Clerical; 4 = Service; 5 = Professional; 6 = Other
SEX		0 = Male; 1 = Female
SOUTH		0 = Does not live in South; 1 = Lives in South
WAGE	Wage (dollars per hour)	