Chapter 2 Descriptive Statistics and Data Analysis

Basic Concepts Review Questions

1. Explain the principal types of descriptive statistics measures that are used for describing data.

Answer:
Descriptive statistics – a collection of quantitative measures and methods of describing data. This includes the measure of central tendency, (mean, median mode and proportion.), the measure of dispersion, (range, variance, standard deviation), the measure of shape (skewness, kurtosis) and frequency distributions and histograms.

2. What are frequency distributions and histograms? What information do they provide?

Answer:
Frequency distribution – a tabular summary that shows the frequency of observations in each of several nonoverlapping classes. Histogram – graphical depiction of a frequency distribution in the form of a column chart. Both frequency distribution and the histogram allow us to visually examine the center, dispersion (variability) and shape of a distribution.

3. Provide some examples of data profiles.

Answer:
Data profiling is an analysis of data to better understand relationships in data, as well as similarities and differences. Data profiles are often expressed as percentiles and quartiles. Percentiles are used on standardized tests used for college or graduate school entrance examinations (SAT, ACT, GMAT, GRE, etc.). Percentiles specify the percentage of other test takers who scored at or below the score of a particular individual.

4. Explain how to compute the relative frequency and cumulative relative frequency.

Answer:
Once the classes (bin, intervals) for the distribution are determined, based on the range of data and the desired number of bins, the relative frequency is computed by counting how many observations fall into each of the bins and then divided by the total number of observations. Cumulative relative frequency – the running total of relative frequencies up to the upper level of each bin.

5. Explain the difference between the mean, median, mode, and midrange. In what situations might one be more useful than the others?

Answer:
Mean – an arithmetic average of a set of observations and is the most appropriate tool for interval and ratio data without significant outliers. Median – the middle point of a sorted set of observations, and is the most appropriate tool for ordinal, interval and ratio data and is not affected by outliers. Mode – the most frequent data point in a set of observations, and is appropriate only for nominal and ordinal data with few frequently occurring observations. Midrange – the average of the largest and smallest observations, and is appropriate when the number of observations is relatively small and is adversely impacted by the presence of outliers.

6. What statistical measures are used for describing dispersion in data? How do they differ from one another?

Answer:
Range – the difference between the largest and the smallest observation, and is extremely sensitive to outliers. Variance – the average of squared deviations for the mean and is also affected by outliers, but not to the same extent as the range. It is expressed in squared units. Standard deviation – the square root of the variance, and represents and average deviation from the mean.

7. Explain the importance of the standard deviation in interpreting and drawing conclusions about risk.

Answer:
When comparing financial investments such as stocks, investors compare average returns, but also risks. If 2 stocks have average returns, and the standard deviation is much higher than the other, than we may conclude that the stock with the higher standard deviation is riskier or more volatile.

8. What does Chebyshev's theorem state and how can it be used in practice?

Answer:
Chebyshev's Theorem – for any set of data, the proportion of values that lie within k standard deviations of the mean is at least $1 - 1/k^2$. In practice, this tells us that for $k = 2$ at least 75% of the observations lie within 2 standard deviations of the mean, and for $k = 3$ at least 89% of the observations lie within 3 standard deviations of the mean.

9. Explain the coefficient of variation and how it can be used.

Answer:
Coefficient of variation – provides a relative measure of the dispersion in data relative to the mean. This allows a researcher to compare 2 stocks that have different means and standard deviations. For the stock with the larger coefficient of variation, we could say that it took more risk per unit of return than the other stock did.

10. Explain the concepts of skewness and kurtosis and what they tell about the distribution of data.

Answer:
Skewness – represents the degree of asymmetry of a distribution around its mean. The closer skewness gets to zero, the closer the distribution is to a perfectly symmetrical one. Positive numbers represent right-skewed distributions, and negative numbers represent a distribution that is left skewed. Kurtosis refers to the peakedness (high and narrow) or flatness of a distribution. The higher the kurtosis, the more area the distribution has in its tails rather than in the middle.

11. Explain the concept of correlation and how to interpret correlation coefficients of 0.3, 0, and –0.95.

Answer:
Correlation – a measure of the strength of a linear relationship between 2 variables. The correlation of 0 implies lack of relationship, correlation of 0.3 represents a weak positive relationship, and a correlation of -0.95 represents a strong negative relationship.

12. What is a proportion? Provide some practical examples where proportions are used in business.

Answer:
Proportion – the fraction of data that have a certain characteristic. It is used mostly with categorical data, such as marketing survey responses. A typical business example might be, “What proportion of school aged children buy a school lunch every day.”

13. What is a cross-tabulation? How can it be used by managers to provide insight about data, particularly for marketing purposes?

Answer:

Cross-tabulation – is a tabular method that displays the number of observations in a data set for different subcategories of two categorical variables, resulting in a contingency table. Managers might look at a contingency table showing total sales by gender and product category, in order to determine which market segment better responds to which product group and adjust their marketing efforts accordingly.

14. Explain the information contained in box plots and dot-scale diagrams.

Answer:
Box plots – graphically display five key statistics of a data set, the minimum, first quartile, median, third quartile, and maximum, and are very useful in identifying the shape of a distribution and outliers in the data. Dot-scale diagrams – shows a histogram of data values as dots corresponding to individual data points, along with the mean, median, first and third quartiles, and ±1, 2, and 3 standard deviation ranges from the mean. The mean acts as a fulcrum as if the data were balanced along an axis.

15. What is a PivotTable? Describe some of the key features that PivotTables have.

Answer:
PivotTables allows you to create custom summaries and charts of key information in the data. PivotTables also provide an easy method of constructing cross - tabulations for categorical data. The beauty of PivotTables is that if you wish to change the analysis, you can simply uncheck the boxes in the PivotTable Field List or drag the variable names to different field areas. You may easily add multiple variables in the fields to create different views of the data.

16. Explain how to compute the mean and variance of a sample and a population. How would you explain the formulas in simple English?

Answer:
If a population consists of *N* observations $x_1, \ldots, x_N$, population mean, μ is calculated as the ratio of sum of the observations $x_1, \ldots, x_N$ to the total number of observations, N. The mean of a sample of n observations, x1, . . . , xn, denoted by "x-bar" is calculated as the ratio of sum of the observations, x1, . . . , xn to the total number of observations, n.
Variance of a population is the sum of the squared deviations of the observations $x_1, \ldots, x_N$ from its mean ,μ divided by the total number of observations, N
Variance of a population is the sum of the squared deviations of the observations $x_1, \ldots, x_n$ from its mean ,x bar divided by the total number of observations minus one.

17. How can one estimate the mean and variance of data that are summarized in a grouped frequency distribution? Why are these only estimates?

Answer:
When data are summarized in a grouped frequency distribution the mean of the data is estimated as = $\frac{\sum_{i=1}^{n} fi\ xi}{n}$ Variance of data is given as .
They are only estimates since they are calculated using the sample data.

18. Explain the concept of covariance. How is covariance used in computing the correlation coefficient?

Answer:
Covariance – Covariance between two (linearly) related variables is the average of the products of deviations of each variable's observation from its respective mean. If, for most of the observations, both variables are either above or below their means at the same time, the covariance will be positive. On the other hand, if for most of the observations, when one variable is above its mean and the other is below its mean, and vice versa, the covariance will be negative. Correlation between the two (linearly) related variables is the covariance, adjusted (divided) by the standard deviations of each of the two variables.

Problems and Applications

1. A community health status survey obtained the following demographic information from the respondents:

Age	Frequency
18-29	297
30-45	661
46-64	634
65+	369

Compute the relative frequency and cumulative relative frequency of the age groups. Also, estimate the average age of the sample of respondents. What assumptions do you have to make to do this?

Answer:

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
18-29	297	15%	15%
30-45	661	34%	49%
46-64	634	32%	81%
65+	369	19%	100%
Total	1961	100%	100%

Assumptions:
1. Assume the distribution within each age category is uniform, so median is the appropriate methodology
2. Use average life expectancy of age 78* for maximum age in 65+ category

Median age/Midpoint	Frequency	Relative Frequency	Weighted Average
23.5	297	15%	3.559153493
37.5	661	34%	12.64023457
55	634	32%	17.78174401
71.5	369	19%	13.45410505
Average age in study	1961	100%	47.43523712

Link used: en.wikipedia.org/wiki/List_of_countries_by_life_expectancy

2. The Excel file *Insurance Survey* provides demographic data and responses to satisfaction with current health insurance and willingness to pay higher premiums for a lower deductible for a sample of employees at a company. Construct frequency distributions and compute the relative frequencies for the categorical variables of gender, education, and marital status. What conclusions can you draw?

Answer:

***Satisfaction

Gender	Frequency	Relative Frequency	Cumulative Relative Frequency
F	9	64%	64%
M	5	36%	100%
Total	14	100%	100%

*** assumes a satisfaction score of 4 or 5 means satisfied

Conclusion, 64% of the satisfied respondents with current insurance are female and 36% of the satisfied insured are male.

Gender	Frequency	Relative Frequency	Cumulative Relative Frequency
F	5	83%	83%
M	1	17%	100%
Total	6	100%	100%

Conclusion, 50% of the respondents who are favorable to new premiums insurance are female and 50% of the respondents who are favorable to new premiums are male.

Gender	Frequency	Relative Frequency	Cumulative Relative Frequency
F	14	58%	58%
M	10	42%	100%
Total	24	100%	100%

58% of the respondents are female and 42% are male

Educational Level	Frequency	Relative Frequency	Cumulative Relative Frequency
College graduate	9	38%	38%
Graduate degree	8	33%	71%
Some college	7	29%	100%
Total	24	100%	100%

38% of respondents are college graduates, 33% have a graduate degree and 29% have some college.

Marital Status	Frequency	Relative Frequency	Cumulative Relative Frequency
Divorced	5	21%	21%
Married	17	71%	92%
Single	1	4%	96%
Widowed	1	4%	100%
Total	24	100%	100%

71% of the respondents are married, 21% are divorced, 4% are single and 4% are widowed.

3. Construct a frequency distribution and histogram for the taxi-in time in the Excel file *Atlanta Airline Data* using the Excel Histogram tool. Use bin ranges from 0 to 50 with widths of 10. Find the relative frequencies and cumulative relative frequencies for each bin, and estimate the average time using the frequency distribution.

Answer:

Flight Number	Origin Airport	Scheduled Arrival Time	Actual Arrival Time	Time Difference (Minutes)	Taxi-in Time (Minutes)
8	IAH	19:04	19:19	15	14
16	LAX	15:10	15:04	-6	6
22	MSY	16:33	16:24	-9	11
24	LAS	14:33	14:27	-6	9
28	MCO	14:10	14:15	5	13
38	MCO	16:10	15:48	-22	6
57	JFK	19:41	19:54	13	12
61	LAX	19:02	19:22	20	11
64	LAS	18:00	17:58	-2	10
66	DFW	15:18	15:14	-4	9
68	SFO	14:44	14:35	-9	7
74	MIA	15:41	15:39	-2	18
101	LAX	17:41	17:56	15	13
105	DTW	17:35	17:26	-9	8
108	MCO	17:09	16:52	-17	11
116	LAX	16:19	16:18	-1	7
130	SLC	14:15	14:38	23	7
147	EWR	19:32	19:19	-13	23
151	SLC	15:25	15:50	25	12
152	LAX	20:31	20:43	12	21
365	LGA	10:53	10:33	-20	9
371	IAD	07:34	07:21	-13	7
373	RDU	08:44	09:09	25	9
377	MSP	13:49	14:12	23	11
409	CLT	08:48	09:17	29	8
418	SJU	11:07	10:59	-8	6
420	SJU	13:05	13:02	-3	11
422	SJU	17:24	17:06	-18	6
424	SJU	18:43	18:22	-21	7
428	SJU	19:40	19:42	2	17
438	STX	19:06	19:06	0	23
509	ROC	08:55	08:26	-29	7
529	CHS	07:22	07:02	-20	11
543	DFW	08:42	09:11	29	19
547	SNA	16:02	15:43	-19	10
660	STT	17:15	17:13	-2	6
665	ORD	09:00	09:02	2	15

674	STT	19:11	19:18	7	12
675	MSP	09:00	10:03	63	13
676	STT	20:34	20:28	-6	9
687	CVG	08:49	08:40	-9	22
1,005	PHL	08:33	09:03	30	7
1,007	PHL	10:04	10:45	41	10
1,009	PHL	11:02	11:09	7	9
1,013	PHL	14:03	14:01	-2	12
1,014	ABQ	13:07	13:02	-5	5
1,015	PHL	15:18	15:10	-8	10
1,016	SAT	14:10	14:19	9	8
1,017	PHL	16:26	16:22	-4	9
1,021	PHL	19:22	19:01	-21	11
1,022	PNS	18:03	17:55	-8	14
1,023	PHL	20:52	20:27	-25	10
1,024	PHX	12:43	12:54	11	9
1,026	PHX	13:49	13:53	4	11
1,030	PHX	17:49	17:40	-9	10
1,032	PHX	19:40	22:18	158	9
1,035	CMH	16:40	16:30	-10	10
1,036	PHX	06:18	06:05	-13	7
1,038	SAN	13:34	13:32	-2	7
1,041	JAX	09:59	09:24	-35	7
1,044	SAN	18:27	18:04	-23	11
1,048	SAN	05:37	05:35	-2	14
1,050	SEA	13:48	13:56	8	12
1,052	SEA	14:57	15:12	15	10
1,054	SEA	19:40	20:02	22	10
1,055	TPA	09:14	09:12	-2	13
1,060	SEA	06:12	06:05	-7	9
1,064	SFO	13:37	13:36	-1	8
1,066	SFO	16:07	16:07	0	9
1,068	SFO	19:40	19:42	2	13
1,070	SFO	21:59	21:49	-10	6
1,074	SFO	06:21	06:07	-14	13
1,077	MSP	11:16	12:36	80	9
1,078	LAS	13:21	13:17	-4	6
1,082	LAS	17:05	17:03	-2	9
1,084	BDL	15:50	15:38	-12	9
1,085	MCO	09:24	09:22	-2	8
1,086	LAS	19:42	19:55	13	29
1,088	LAS	20:37	20:25	-12	12
1,091	CMH	14:09	14:10	1	10
1,092	LAS	06:13	06:02	-11	9
1,118	MCO	18:34	18:31	-3	15
1,122	EYW	14:45	14:42	-3	10
1,136	RSW	18:07	17:48	-19	13
1,140	PBI	20:49	20:55	6	39

1,148	MCO	13:34	13:33	-1	13
1,159	BUF	18:59	18:36	-23	12
1,162	PBI	16:40	17:24	44	13
1,164	DTW	09:00	08:42	-18	12
1,175	DTW	12:37	12:46	9	7
1,177	RDU	13:53	13:47	-6	10
1,186	JAX	15:00	14:45	-15	16
1,202	RSW	12:44	12:39	-5	11
1,213	ROC	18:10	18:06	-4	12
1,215	SRQ	15:03	14:54	-9	10
1,221	BNA	08:57	08:51	-6	9
1,228	RSW	15:22	15:17	-5	17
1,248	PBI	12:59	13:05	6	11
1,253	PIT	09:00	08:32	-28	11
1,258	MSY	08:34	08:43	9	17
1,259	RDU	18:44	19:15	31	49
1,270	MCI	15:57	16:36	39	23
1,271	RSW	13:50	13:28	-22	8
1,279	BDL	13:18	13:01	-17	8
1,291	STL	15:12	15:02	-10	16
1,292	PBI	10:05	09:55	-10	17
1,296	TUS	19:25	19:14	-11	14
1,297	JFK	08:49	08:39	-10	14
1,302	MCO	06:57	06:47	-10	7
1,304	MCO	08:11	07:52	-19	8
1,306	FLL	18:56	19:04	8	13
1,308	MCO	10:19	10:16	-3	16
1,310	MCO	11:04	10:50	-14	8
1,312	MCO	12:05	11:54	-11	8
1,314	MCO	13:08	13:07	-1	10
1,318	MCO	15:10	14:57	-13	15
1,324	MCO	18:10	18:04	-6	14
1,326	MCO	19:16	19:00	-16	15
1,328	MCO	20:14	19:53	-21	9
1,483	PIT	10:10	10:57	47	7
1,494	SAT	08:48	08:48	0	10
1,500	JAX	18:30	18:29	-1	9
1,502	SAT	11:10	11:05	-5	9
1,510	MEM	13:59	14:00	1	7
1,512	SLC	16:40	16:45	5	16
1,513	CMH	07:45	07:26	-19	8
1,516	AUS	19:17	19:31	14	16
1,517	JAX	06:40	06:31	-9	7
1,518	SNA	14:14	14:06	-8	8
1,520	JAX	17:05	17:02	-3	8
1,521	PHL	20:01	20:04	3	14
1,528	PBI	18:09	17:57	-12	12
1,531	JAX	11:35	11:17	-18	11

1,536	JAC	19:07	19:30	23	45
1,538	PDX	14:06	14:24	18	9
1,542	SLC	19:28	19:41	13	18
1,553	SAV	16:31	16:14	-17	10
1,554	BHM	08:11	08:30	19	30
1,555	BUF	16:28	16:13	-15	9
1,559	BDL	09:56	09:37	-19	15
1,561	JAX	07:44	07:32	-12	9
1,563	MSP	18:08	19:10	62	10
1,564	IND	19:12	19:15	3	18
1,565	MSP	14:55	15:06	11	10
1,577	MCI	10:10	09:54	-16	8
1,586	SLC	12:38	12:51	13	13
1,588	PBI	19:40	19:43	3	19
1,591	RDU	12:30	12:19	-11	7
1,598	SNA	18:38	18:30	-8	10
1,599	IAD	10:09	10:02	-7	24
1,601	BDL	09:00	08:34	-26	10
1,604	MDW	18:15	18:18	3	14
1,605	CVG	18:19	18:20	1	29
1,606	MSY	17:42	17:44	2	14
1,610	MCI	08:41	08:49	8	28
1,612	RSW	07:45	07:35	-10	8
1,615	RDU	16:35	16:38	3	13
1,617	MEM	16:09	16:36	27	13
1,618	CHS	16:51	16:39	-12	9
1,620	MSP	19:45	19:58	13	10
1,623	CHS	08:26	08:13	-13	11
1,627	MSP	16:43	16:59	16	14
1,628	MCI	19:09	19:49	40	14
1,629	RDU	09:51	09:51	0	15
1,632	TPA	17:45	17:28	-17	12
1,633	MSY	09:49	09:34	-15	7
1,634	EGE	17:59	17:48	-11	11
1,636	MEM	08:21	08:13	-8	9
1,637	IAD	14:10	14:02	-8	11
1,638	PBI	09:00	09:08	8	9
1,640	MOB	08:23	08:30	7	15
1,641	JFK	11:13	10:55	-18	9
1,649	ORF	08:58	09:04	6	14
1,652	SMF	19:29	19:45	16	17
1,653	MKE	08:53	09:00	7	10
1,655	MSP	21:00	21:19	19	10
1,659	SAV	12:53	12:57	4	7
1,664	MCI	12:34	12:41	7	12
1,675	ABQ	18:24	18:27	3	10
1,684	SJC	13:57	14:18	21	11
1,688	RSW	19:34	19:28	-6	18

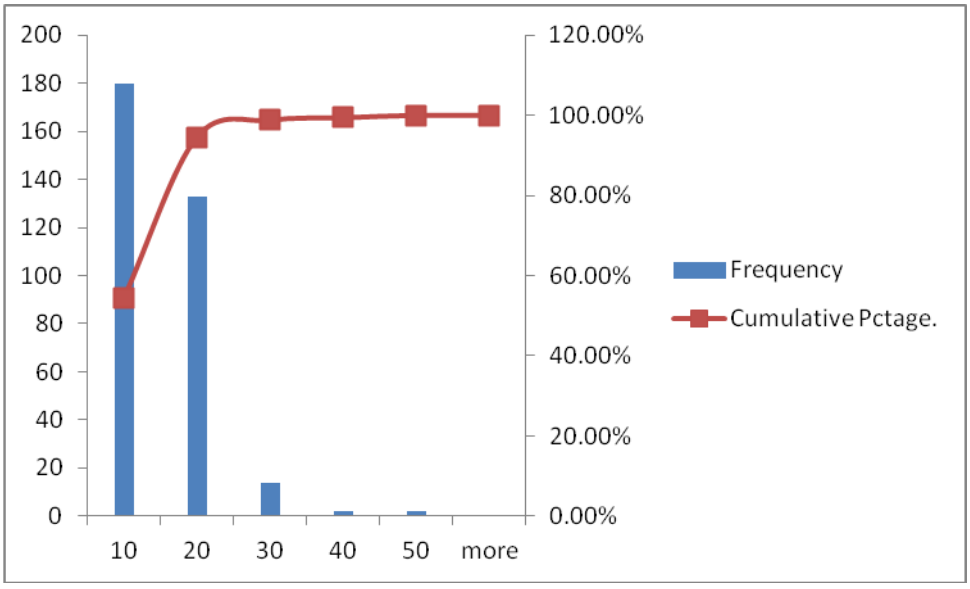
1,689	BDL	18:10	17:51	-19	9
1,693	SNA	21:20	21:00	-20	7
1,694	SLC	22:37	22:49	12	8
1,695	CMH	18:45	18:54	9	16
1,696	STL	08:45	08:58	13	10
1,703	CMH	08:57	09:00	3	11
1,705	DTW	20:08	20:22	14	8
1,706	AUS	10:01	10:02	1	14
1,708	DTW	07:45	07:54	9	8
1,709	ORD	10:10	11:09	59	9
1,711	DTW	11:26	11:27	1	11
1,714	ONT	13:41	14:07	26	12
1,716	CLT	09:43	10:02	19	12
1,717	IND	08:56	08:55	-1	15
1,720	SLC	06:20	06:40	20	8
1,723	JFK	13:46	13:17	-29	7
1,727	JFK	22:04	21:19	-45	8
1,728	RIC	07:45	07:24	-21	8
1,731	DAY	07:45	07:39	-6	17
1,734	MIA	20:53	20:45	-8	13
1,737	JFK	16:40	16:10	-30	8
1,738	SRQ	12:49	12:39	-10	12
1,739	CVG	16:16	16:14	-2	9
1,740	MSY	12:29	12:42	13	11
1,747	MSP	10:08	10:54	46	15
1,759	MDW	10:05	09:56	-9	13
1,766	HDN	18:04	18:03	-1	11
1,769	LGA	08:44	08:19	-25	9
1,771	LGA	09:46	09:39	-7	9
1,775	LGA	12:02	11:24	-38	10
1,779	LGA	13:58	13:28	-30	7
1,781	LGA	14:53	14:18	-35	11
1,783	LGA	15:50	15:32	-18	9
1,785	LGA	16:50	16:27	-23	8
1,787	LGA	17:55	17:39	-16	28
1,789	LGA	18:54	18:37	-17	15
1,790	SAT	16:40	17:15	35	10
1,793	LGA	20:55	20:28	-27	9
1,797	LGA	22:49	22:22	-27	9
1,844	SLC	21:02	20:54	-8	10
1,850	SAV	06:41	06:33	-8	7
1,851	BOS	08:52	08:25	-27	9
1,852	SRQ	07:39	07:36	-3	10
1,853	BOS	10:21	10:07	-14	20
1,854	PNS	09:06	09:08	2	14
1,855	BOS	11:42	11:13	-29	8
1,857	BOS	12:38	12:00	-38	5
1,859	BOS	14:07	13:48	-19	9

1,861	BOS	15:25	14:55	-30	9
1,865	BOS	17:49	17:25	-24	9
1,867	BOS	18:51	19:04	13	35
1,869	BOS	19:58	20:06	8	9
1,877	BWI	08:50	08:52	2	20
1,878	SAV	09:00	09:15	15	9
1,879	BWI	10:05	10:04	-1	11
1,881	BWI	11:57	11:56	-1	14
1,882	SAV	07:42	07:49	7	11
1,883	BWI	13:08	14:22	74	10
1,884	TUS	12:33	12:32	-1	9
1,885	BWI	14:33	14:36	3	9
1,887	BWI	15:35	15:25	-10	8
1,889	BWI	18:09	17:53	-16	7
1,891	PDX	19:39	19:50	11	9
1,896	DEN	05:56	06:17	21	6
1,897	PBI	07:44	07:51	7	10
1,898	DEN	11:09	11:48	39	7
1,899	RIC	08:51	13:16	265	7
1,900	DEN	12:22	13:00	38	7
1,902	DEN	13:36	14:29	53	9
1,904	DEN	15:59	15:59	0	8
1,908	DEN	18:10	18:44	34	7
1,910	DEN	20:45	20:55	10	11
1,914	DFW	10:08	10:13	5	9
1,917	BUF	08:48	08:37	-11	9
1,918	DFW	12:45	12:52	7	10
1,920	DFW	14:03	14:35	32	10
1,921	CHS	13:09	13:19	10	10
1,924	DFW	16:30	04:23	713	8
1,926	DFW	19:15	02:44	449	13
1,935	ORF	11:49	11:33	-16	9
1,943	ORD	15:10	15:08	-2	15
1,945	ORD	18:20	18:34	14	10
1,948	SAN	15:03	14:58	-5	8
1,951	DCA	08:00	07:59	-1	9
1,953	DCA	09:17	09:06	-11	14
1,954	DAB	07:30	07:38	8	11
1,955	DCA	10:05	10:01	-4	12
1,959	DCA	12:02	11:48	-14	11
1,960	PNS	10:10	10:03	-7	16
1,961	DCA	12:58	12:54	-4	8
1,962	ABQ	11:16	11:30	14	8
1,964	COS	12:33	12:46	13	11
1,965	DCA	15:01	14:59	-2	11
1,967	DCA	16:05	16:07	2	13
1,969	DCA	17:03	16:54	-9	8
1,971	DCA	18:00	18:08	8	11

1,973	DCA	19:09	19:05	-4	13
1,975	DCA	20:05	19:55	-10	10
1,978	SAT	19:26	19:25	-1	21
1,982	MIA	08:35	09:02	27	16
1,984	MIA	09:49	09:41	-8	12
1,988	MIA	13:28	13:36	8	11
1,989	IND	10:08	09:51	-17	11
1,990	MIA	14:30	14:32	2	23
1,991	EWR	09:00	08:49	-11	12
1,992	RSW	08:47	08:51	4	20
1,994	MIA	16:55	16:46	-9	14
1,995	BUF	13:54	13:53	-1	8
1,996	MIA	18:10	18:23	13	17
1,998	MIA	19:37	19:25	-12	10
1,999	JAX	09:00	09:19	19	9
2,007	EWR	10:10	10:03	-7	18
2,008	SRQ	08:44	08:52	8	14
2,009	EWR	11:13	15:33	260	12
2,011	EWR	13:04	12:45	-19	8
2,014	ELP	12:44	12:59	15	8
2,015	EWR	15:25	15:05	-20	9
2,016	JAX	13:09	15:35	146	8
2,017	EWR	16:39	16:05	-34	7
2,019	EWR	18:01	17:44	-17	11
2,028	FLL	08:50	08:49	-1	14
2,030	FLL	09:58	10:12	14	15
2,032	FLL	10:49	10:41	-8	10
2,034	FLL	12:07	12:01	-6	6
2,036	FLL	13:39	13:27	-12	9
2,042	FLL	15:54	15:45	-9	6
2,044	FLL	17:03	16:52	-11	8
2,046	FLL	18:26	18:06	-20	10
2,048	FLL	19:39	19:30	-9	13
2,050	FLL	20:52	20:48	-4	20
2,054	TPA	07:03	06:51	-12	7
2,056	TPA	08:09	07:58	-11	5
2,060	TPA	10:10	10:11	1	18
2,062	TPA	11:25	11:20	-5	9
2,064	TPA	12:46	12:47	1	13
2,066	TPA	14:04	13:49	-15	12
2,068	TPA	16:14	16:05	-9	10
2,072	TPA	18:59	19:00	1	29
2,074	TPA	20:14	19:49	-25	10
2,076	MSY	15:02	14:42	-20	7
2,079	MDW	12:40	12:52	12	8
2,080	LAX	13:24	13:20	-4	7
2,085	JAX	10:45	10:32	-13	9
2,086	PBI	14:52	14:29	-23	8

2,088	IAH	12:42	13:10	28	12
2,092	LAX	21:55	22:05	10	11
2,094	LAX	23:58	23:36	-22	7
2,096	LAX	06:10	05:49	-21	8
2,097	CLT	10:37	10:28	-9	9
2,098	LAX	07:08	06:56	-12	8

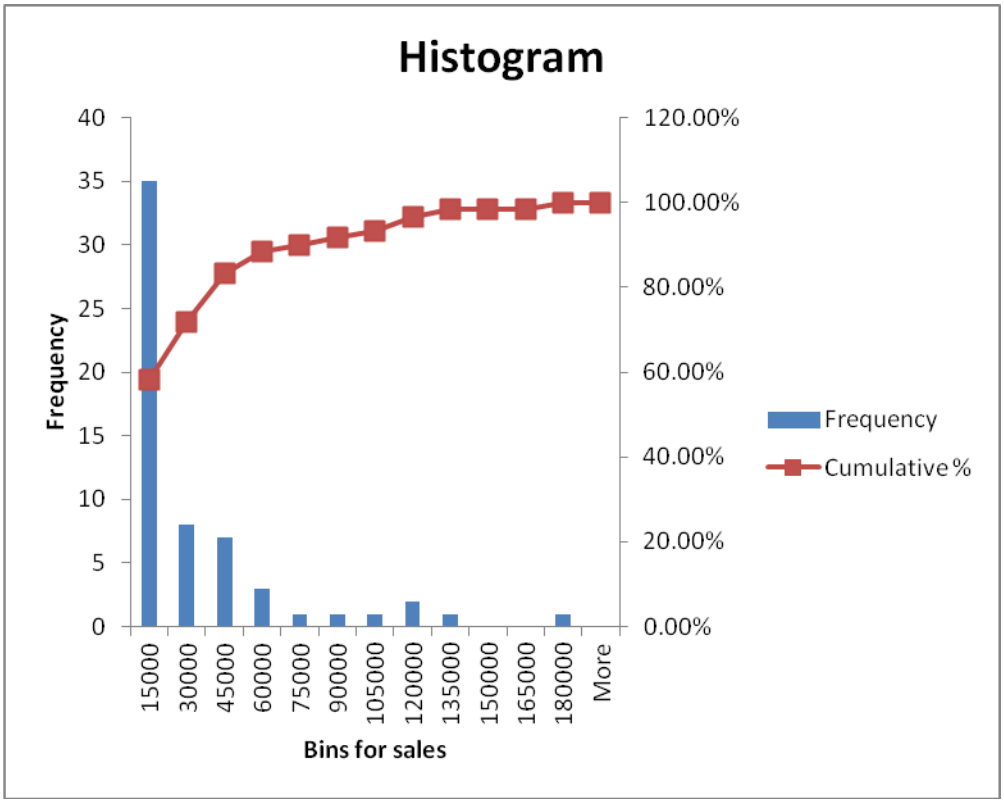
Bins for Taxi-in-time	*Frequency*	*Cumulative %*
10	180	54.38%
20	133	94.56%
30	14	98.79%
40	2	99.40%
50	2	100.00%
More	0	100.00%



4. Construct frequency distributions and histograms using the Excel Histogram tool for the *Gross Sales and Gross Profit data* in the Excel file Sales Data. Define appropriate bin ranges for each variable.

Answer:

Bins for sales	*Frequency*	*Cumulative %*
15000	35	58.33%
30000	8	71.67%
45000	7	83.33%
60000	3	88.33%
75000	1	90.00%
90000	1	91.67%
105000	1	93.33%
120000	2	96.67%
135000	1	98.33%
150000	0	98.33%
165000	0	98.33%
180000	1	100.00%
More	0	100.00%



5. Find the 10th and 90th percentiles of home prices in the Excel file *Home Market Value*.

Answer:

Home market value	**Prices**
90th percentile	$108,090.00
10th percentile	$81,320.00

6. Find the first, second, and third quartiles for each of the performance statistics in the Excel file *Ohio Education Performance*. What is the interquartile range for each of these?

Answer:

	Writing	**Reading**	**Math**	**Citizenship**	**Science**	**All**
First Quartile	82	75.5	52	68.5	62.5	40
Second Quartile	87	83	66	78	75	52
Third Quartile	91	88	73.5	84.5	82.5	64
Interquartile range	9	12.5	21.5	16	20	24

7. Find the 10th and 90th percentiles and the first and third quartiles for the time difference between the scheduled and actual arrival times in the *Atlanta Airline Data* Excel file.

Answer:

	Time difference between scheduled and actual	
First Quartile	-12	min
Third Quartile	8	min
	a negative value indicates early arrival	
10th Percentile	-20	min
90th Percentile	23	min

8. Compute the mean, median, variance, and standard deviation using the appropriate Excel functions for all the variables in the Excel file *National Football League*. Note that the data represent a population. Apply the Descriptive Statistics tool to these data, what differences do you observe? Why did this occur?

Answer:

	Mean	Population Variance	Sample Variance	Pop Std Deviation	Sample Std Deviation
Points/Game	21.69375	24.14433594	24.92318548	4.913688628	4.992312639
Yards/Game	325.21875	1218.714648	1258.028024	34.91009379	35.46869076
Rushing Yards/Game	110.9125	382.3692187	394.7037097	19.55426344	19.86715152
Passing Yards/Game	214.30938	1274.4596	1315.5712	35.69957422	36.27080368
Opponent Yards/Game	325.23125	706.1390234	728.9177016	26.57327649	26.99847591
Opponent Rushing Yards/Game	110.93125	344.3908984	355.5002823	18.55777191	18.85471512
Opponent Passing Yards/Game	214.32188	508.223584	524.6178931	22.54381476	22.9045387
Penalties	91.625	293.609375	303.0806452	17.13503356	17.4092115
Penalty Yards	720.0625	20735.93359	21404.83468	143.9997694	146.303912
Interceptions	16.6875	14.46484375	14.93145161	3.80326751	3.864123654
Fumbles	12	12.9375	13.35483871	3.596873642	3.654427275
Passes Intercepted	16.6875	19.33984375	19.96370968	4.397708921	4.468076731
Fumbles Recovered	12	19.75	20.38709677	4.444097209	4.515207279

Absolute Difference

	Sample - Pop Variance	Sample - Pop Variance	Sample - Pop Std Dev
Points/Game	0.778849546	0.778849546	0.07862401
Yards/Game	39.31337576	39.31337576	0.558596969
Rushing Yards/Game	12.33449093	12.33449093	0.312888082
Passing Yards/Game	41.11159999	41.11159999	0.571229458
Opponent Yards/Game	22.77867818	22.77867818	0.425199422
Opponent Rushing Yards/Game	11.10938382	11.10938382	0.296943205
Opponent Passing Yards/Game	16.39430916	16.39430916	0.360723941
Penalties	9.471270161	9.471270161	0.274177946
Penalty Yards	668.9010837	668.9010837	2.304142615
Interceptions	0.466607863	0.466607863	0.060856144
Fumbles	0.41733871	0.41733871	0.057553633
Passes Intercepted	0.623865927	0.623865927	0.070367811
Fumbles Recovered	0.637096774	0.637096774	0.071110071

Relative Difference

	Sample/Pop Variance	Sample/Pop Std Dev
Points/Game	1.032258065	1.016001016
Yards/Game	1.032258065	1.016001016
Rushing Yards/Game	1.032258065	1.016001016
Passing Yards/Game	1.032258065	1.016001016
Opponent Yards/Game	1.032258065	1.016001016
Opponent Rushing Yards/Game	1.032258065	1.016001016
Opponent Passing Yards/Game	1.032258065	1.016001016
Penalties	1.032258065	1.016001016
Penalty Yards	1.032258065	1.016001016
Interceptions	1.032258065	1.016001016
Fumbles	1.032258065	1.016001016
Passes Intercepted	1.032258065	1.016001016
Fumbles Recovered	1.032258065	1.016001016

From the above table we can observe that the sample variance is about 3% higher than the population variance. Sample standard deviation is about 2% higher than the population standard deviation. The difference occurs due to the different denominators used to average the squared deviations from the mean for populations and samples.

9. Data obtained from a county auditor in the Excel file *Home Market Value* provides information about the age, square footage, and current market value of houses along one street in a particular subdivision.
a. Considering these data as a sample of homeowners on this street, compute the mean, variance, and standard deviation for each of these variables using the formulas (2A.2), (2A.5), and (2A.7).
b. Compute the coefficient of variation for each variable. Which has the least and greatest relative dispersion?

Answer:

a.

	House Age	**Square Feet**	**Market Value**
Mean	29.83	1695.26	$92,069.05
Median	28	1666	$88,500.00
Variance	5.76	47357.96	$108715946.71
Standard Deviation	2.40	217.62	$10,426.69
Coefficient of Variation	0.08	0.13	0.11

b. Higher the Coefficients of variation higher greatest is the relative dispersion and vice versa

Coefficients of variation indicate that square footage has the highest dispersion and age the lowest dispersion around the respective means.

10. The Excel file *Seattle Weather* contains weather data for Seattle, Oregon. Apply the Descriptive Statistics tool to these data. Show that Chebyshev's theorem holds for the average temperature and rainfall.

Answer:

Temperature		*Rainfall*		*Clear*	
Mean	52.775	Mean	3.175	Mean	5.916667
Standard Error	2.594257207	Standard Error	0.521671	Standard Error	0.891529
Median	51.95	Median	2.9	Median	5
Mode	#N/A	Mode	#N/A	Mode	3
Standard Deviation	8.98677058	Standard Deviation	1.80712	Standard Deviation	3.088346
Sample Variance	80.76204545	Sample Variance	3.265682	Sample Variance	9.537879
Kurtosis	-1.529095045	Kurtosis	-1.314	Kurtosis	-0.49101
Skewness	0.193004576	Skewness	0.408552	Skewness	0.804531
Range	24.4	Range	5.1	Range	9
Minimum	41.3	Minimum	0.9	Minimum	3
Maximum	65.7	Maximum	6	Maximum	12
Sum	633.3	Sum	38.1	Sum	71
Count	12	Count	12	Count	12

Partly Cloudy		*Cloudy*	
Mean	7.75	Mean	16.75
Standard Error	0.538305	Standard Error	1.309204
Median	8	Median	17
Mode	8	Mode	23
Standard Deviation	1.864745	Standard Deviation	4.535216
Sample Variance	3.477273	Sample Variance	20.56818
Kurtosis	-1.26423	Kurtosis	-0.90566
Skewness	-0.27129	Skewness	-0.16972
Range	5	Range	14
Minimum	5	Minimum	9
Maximum	10	Maximum	23
Sum	93	Sum	201
Count	12	Count	12

Chebyshev k	k*s	k*s	1-1/k^2
2	17.97	3.61	75%
3	26.96	5.42	89%

	Temperature	*Rainfall*
$\bar{x}$ – 2s	34.81	-0.44
$\bar{x}$ + 2s	70.75	6.79

	Temperature	*Rainfall*
$\bar{x}$ – 3s	25.82	-2.25
$\bar{x}$ + 3s	79.74	8.60

Actual observations	*Temperature*	*Rainfall*	% within
within 2 s	12	12	100%
within 3 s	12	12	100%

11. The Excel file Baseball Attendance shows the attendance in thousands at San Francisco Giants baseball games for the 10 years before the Oakland A's moved to the Bay Area in 1968, as well as the combined attendance for both teams for the next 11 years. What is the mean and standard deviation of the number of baseball fans attending before and after the A's move to the San Francisco area? What conclusions might you draw?

Answer:

	San Francisco Giants (1958-1967 seasons)	Giants + Oakland A's after move to Oakland (1968-1978)
Attendance		
Average	1499.4	1646.2
Std Deviation	171.12	304.0

The average attendance only increased by about 150 fans after the move to Oakland. The variability, however, nearly doubled. The primary reason was the higher attendance in 1971 and 1978.

12. For the Excel file University Grant Proposals, compute descriptive statistics for all proposals and also for the proposals that were funded and those that were rejected. Are any differences apparent?

Answer:

All Projects $		*Funded Project $*		*Rejected Project $*	
Mean	643382.5102	Mean	378563.0741	Mean	866250.3525
Standard Error	33018.54871	Standard Error	37523.54823	Standard Error	50934.4249
Median	250195	Median	144508.5	Median	383750
Mode	1918750	Mode	25000	Mode	1918750
Standard Deviation	1424014.635	Standard Deviation	1093990.023	Standard Deviation	1618721.346
Sample Variance	2.02782E+12	Sample Variance	1.19681E+12	Sample Variance	2.62026E+12
Kurtosis	84.96076528	Kurtosis	160.0853178	Kurtosis	63.70798877
Skewness	7.811457769	Skewness	11.01485026	Skewness	6.723383576
Range	21058500	Range	20139887	Range	21058100
Minimum	1000	Minimum	1000	Minimum	1400
Maximum	21059500	Maximum	20140887	Maximum	21059500
Sum	1196691469	Sum	321778613	Sum	874912856
Count	1860	Count	850	Count	1010

Acceptance rate	46%	Rejection rate	54%
Difference between funded and rejected projects, based on means			$487,687
Difference between funded and rejected projects, based on medians			$13,411
Difference between funded and rejected projects, based on modes			$239,242

Measures of shape
Both the reject and funded projects had distributions than were skewed strongly to the right. Funded projects were skewed more to the right. This means that most of the funded projects were requesting a comparatively lower amount. The funded projects had a high kurtosis. The bulk of the rejected proposals ranged from $120,000 to 1,200,000, also indicated by a lower kurtosis.

Measures of Central Tendency
Lightly more than half of the projects were rejected
Rejected projects were higher, on average, by almost $500,000.
The mean/median/mode were most different on rejected projects.

Both ranges and standard deviations for funded and rejected projects are comparable. The rejected projects had much higher dispersions.

Funded Project $	
Mean	378563.0741
Standard Error	37523.54823
Median	144508.5
Mode	25000
Standard Deviation	1093990.023
Sample Variance	1.19681E+12
Kurtosis	160.0853178
Skewness	11.01485026
Range	20139887
Minimum	1000
Maximum	20140887
Sum	321778613
Count	850

Rejected Project $	
Mean	866250.3525
Standard Error	50934.4249
Median	383750
Mode	1918750
Standard Deviation	1618721.346
Sample Variance	2.62026E+12
Kurtosis	63.70798877
Skewness	6.723383576
Range	21058100
Minimum	1400
Maximum	21059500
Sum	874912856
Count	1010

13. Compute descriptive statistics for liberal arts colleges and research universities in the Excel file *Colleges and Universities*. Compare the two types of colleges. What can you conclude?

Answer:

Liberal arts colleges

Median SAT		*Acceptance Rate*		*Expenditures/Student*	
Mean	1256.64	Mean	0.4056	Mean	21611.56
Standard Error	8.734773418	Standard Error	0.025033844	Standard Error	724.8700092
Median	1255	Median	0.38	Median	20377
Mode	1300	Mode	0.36	Mode	#N/A
Standard Deviation	43.67386709	Standard Deviation	0.125169219	Standard Deviation	3624.350046
Sample Variance	1907.406667	Sample Variance	0.015667333	Sample Variance	13135913.26
Kurtosis	-0.732360369	Kurtosis	-0.716038133	Kurtosis	-1.231422743
Skewness	-0.028178376	Skewness	0.392942079	Skewness	0.305870628
Range	166	Range	0.45	Range	11975
Minimum	1170	Minimum	0.22	Minimum	15904
Maximum	1336	Maximum	0.67	Maximum	27879
Sum	31416	Sum	10.14	Sum	540289
Count	25	Count	25	Count	25

Top 10% HS		*Graduation %*	
Mean	67.24	Mean	84.12
Standard Error	2.160462913	Standard Error	1.21836
Median	68	Median	85
Mode	65	Mode	80
Standard Deviation	10.80231457	Standard Deviation	6.091798
Sample Variance	116.69	Sample Variance	37.11
Kurtosis	-0.809206641	Kurtosis	-0.5741
Skewness	-0.14381653	Skewness	-0.45791
Range	39	Range	21
Minimum	47	Minimum	72
Maximum	86	Maximum	93
Sum	1681	Sum	2103
Count	25	Count	25

Research universities

Top 10% HS		*Graduation %*	
Mean	81.45833333	Mean	82.33333

Median SAT		*Acceptance Rate*		*Expenditures/Student*	
Mean	1269.833333	Mean	0.355416667	Mean	38861.125
Standard Error	15.96255431	Standard Error	0.02859626	Standard Error	3690.657672
Median	1280	Median	0.315	Median	37867
Mode	1225	Mode	0.24	Mode	#N/A
Standard Deviation	78.2002261	Standard Deviation	0.14009249	Standard Deviation	18080.45622
Sample Variance	6115.275362	Sample Variance	0.019625906	Sample Variance	326902897.2
Kurtosis	-0.622486887	Kurtosis	-0.780709719	Kurtosis	5.639882783
Skewness	-0.333429842	Skewness	0.503041403	Skewness	1.943611514
Range	291	Range	0.47	Range	82897
Minimum	1109	Minimum	0.17	Minimum	19365
Maximum	1400	Maximum	0.64	Maximum	102262
Sum	30476	Sum	8.53	Sum	932667
Count	24	Count	24	Count	24

Standard Error	2.531668384	Standard Error	1.772032
Median	83.5	Median	86
Mode	95	Mode	90
Standard Deviation	12.40259148	Standard Deviation	8.681147
Sample Variance	153.8242754	Sample Variance	75.36232
Kurtosis	0.814604975	Kurtosis	-0.19671
Skewness	-0.980976299	Skewness	-0.77024
Range	46	Range	32
Minimum	52	Minimum	61
Maximum	98	Maximum	93
Sum	1955	Sum	1976
Count	24	Count	24

Measures of Central Tendency

Median SAT- Both the means and the median are higher for the universities.
Acceptance Rate - Acceptance rates are higher at liberal arts colleges, using mean, median and mode.
Expenditures/Student - Expenditures at liberal arts colleges are much higher, $17, 250 for means, and $17,490 for medians.
Top 10% HS - Both mean and median data show that universities have about 15% higher percentage of the top HS students.
Graduation % - The mean and median show graduation rates to be about the same, with most frequent graduation % to be 10 points higher.

Measures of Dispersion

When examining the data between universities and liberal arts colleges, the major difference is in the expenditure per student. The standard deviation and range for the universities are 5 and 7 times as large, respectively. This indicates that there is a great deal of difference in expenditure of students in universities, compared to liberal arts colleges. For all other variables, universities are slightly more dispersed, but comparable.

Measures of Shape

Distribution of university expenditures is strongly skewed to the right with high peakness, kurtosis as opposed to liberal arts expenditures which have very little positive skewness and actually negative kurtosis, indicating a flatter distribution. The top 10% High School distribution has a strong left skewness, for research universities, the kurtosis opposite to the liberal arts distribution.

14. Compute descriptive statistics for all colleges and branch campuses for each year in the Excel file *Freshman College Data*. Are any differences apparent from year to year?

Answer:

2007

Avg ACT		*Avg SAT*		*HS GPA*	
Mean	22.19091	Mean	1043.036364	Mean	3.108905455
Standard Error	0.738851	Standard Error	30.92917209	Standard Error	0.091212149
Median	21.6	Median	1025.4	Median	3.084
Mode	#N/A	Mode	#N/A	Mode	#N/A
Standard Deviation	2.450492	Standard Deviation	102.5804589	Standard Deviation	0.302516475
Sample Variance	6.004909	Sample Variance	10522.75055	Sample Variance	0.091516218
Kurtosis	-0.18901	Kurtosis	-0.688310488	Kurtosis	-0.677827036
Skewness	0.760534	Skewness	0.493235313	Skewness	-0.060373476

% top 10%		*% top 20%*		*1st year retention rate*	
Mean	0.141909091	Mean	0.303354545	Mean	0.704872665
Standard Error	0.036010467	Standard Error	0.06000226	Standard Error	0.026726439
Median	0.107	Median	0.197	Median	0.701754386
Mode	#N/A	Mode	#N/A	Mode	0.598784195
Standard Deviation	0.119433207	Standard Deviation	0.199004982	Standard Deviation	0.08864157
Sample Variance	0.014264291	Sample Variance	0.039602983	Sample Variance	0.007857328
Kurtosis	-0.288939123	Kurtosis	0.008968745	Kurtosis	2.420733092
Skewness	0.760534	Skewness	0.493235313	Skewness	-0.060373476

2008

Avg ACT		*Avg SAT*		*HS GPA*	
Mean	22.31997	Mean	1038.581909	Mean	3.179309809
Standard Error	0.733369	Standard Error	30.57543801	Standard Error	0.102845733
Median	21.901	Median	1013.24	Median	3.187714
Mode	#N/A	Mode	#N/A	Mode	#N/A
Standard Deviation	2.432309	Standard Deviation	101.4072557	Standard Deviation	0.341100708
Sample Variance	5.916126	Sample Variance	10283.43151	Sample Variance	0.116349693
Kurtosis	-0.34397	Kurtosis	-0.904942935	Kurtosis	-0.256515339
Skewness	0.675573	Skewness	0.415550737	Skewness	-0.667917195

% top 10%		*% top 20%*		*1st year retention rate*	
Mean	0.173654545	Mean	0.329536364	Mean	0.720545455
Standard Error	0.040146159	Standard Error	0.057456154	Standard Error	0.03277842
Median	0.1463	Median	0.3057	Median	0.677
Mode	#N/A	Mode	#N/A	Mode	#N/A
Standard Deviation	0.133149746	Standard Deviation	0.190560506	Standard Deviation	0.108713719
Sample Variance	0.017728855	Sample Variance	0.036313307	Sample Variance	0.011818673
Kurtosis	0.263622105	Kurtosis	-0.54898285	Kurtosis	-0.31698497
Skewness	1.039847456	Skewness	0.652046013	Skewness	0.004502078

2009

Avg ACT		*Avg SAT*		*HS GPA*	
Mean	22.38182	Mean	1047.727273	Mean	3.184454545
Standard Error	0.722221	Standard Error	29.05124028	Standard Error	0.101027752
Median	22	Median	1015.5	Median	3.199
Mode	22.7	Mode	#N/A	Mode	3.199
Standard Deviation	2.395336	Standard Deviation	96.35206371	Standard Deviation	0.335071146
Sample Variance	5.737636	Sample Variance	9283.720182	Sample Variance	0.112272673
Kurtosis	-0.37508	Kurtosis	-0.782277897	Kurtosis	-0.765970291
Skewness	0.286869	Skewness	0.508936698	Skewness	-0.504897576

% top 10%		*% top 20%*		*1st year retention rate*	
Mean	0.164090909	Mean	0.322909091	Mean	0.765090909
Standard Error	0.037545719	Standard Error	0.061156202	Standard Error	0.023384321
Median	0.127	Median	0.275	Median	0.759
Mode	#N/A	Mode	#N/A	Mode	#N/A
Standard Deviation	0.124525061	Standard Deviation	0.202832174	Standard Deviation	0.077557017
Sample Variance	0.015506491	Sample Variance	0.041140891	Sample Variance	0.006015091
Kurtosis	-1.158449554	Kurtosis	-0.81760587	Kurtosis	-0.93286083
Skewness	0.683447072	Skewness	0.634333939	Skewness	0.198054477

2010

Avg ACT		*Avg SAT*		*HS GPA*	
Mean	22.28645	Mean	1046.545297	Mean	3.151636364
Standard Error	0.817044	Standard Error	31.324972	Standard Error	0.114370647
Median	22.01	Median	1019.423	Median	3.2367
Mode	#N/A	Mode	#N/A	Mode	#N/A
Standard Deviation	2.709829	Standard Deviation	103.8931787	Standard Deviation	0.379324524
Sample Variance	7.343171	Sample Variance	10793.79258	Sample Variance	0.143887095
Kurtosis	-1.17914	Kurtosis	-1.207598944	Kurtosis	-1.105359757
Skewness	0.41064	Skewness	0.521615172	Skewness	-0.520242543

% top 10%		*% top 20%*		*1st year retention rate*	
Mean	0.150318182	Mean	0.316490909	Mean	0.737727273
Standard Error	0.03608395	Standard Error	0.059012882	Standard Error	0.026743934
Median	0.1165	Median	0.2949	Median	0.725
Mode	#N/A	Mode	#N/A	Mode	#N/A
Standard Deviation	0.119676922	Standard Deviation	0.195723588	Standard Deviation	0.088699595
Sample Variance	0.014322566	Sample Variance	0.038307723	Sample Variance	0.007867618
Kurtosis	-1.194742506	Kurtosis	-1.20528543	Kurtosis	-1.20160546
Skewness	0.651170062	Skewness	0.318248304	Skewness	0.313499252

There is very slight skewness, except for 1st year retention rate and % top 10 in 2007 and 2008. There is also very little peakness, positive kurtosis, with distributions generally getting flatter over the years.
Means and medians are fairly comparable due to lack of significant skewness, which allows us to use the mean as a measure of the center, and to use standard deviation to measure dispersion.
Most of the measures experienced increases from 2007 - 2009. There was a slight decrease in ACT, SAT and GAP values in 2010.

15. The data in the Excel file Church Contributions were reported on annual giving for a church. Estimate the mean and standard deviation of the annual contributions, using formulas (2A.8) and (2A.10), assuming these data represent the entire population of parishioners.

Answer:

			FrequencyAll Parishoners	**Midpoint**	**Midpoint*Frequency**	**Frq*(Mid-mean)^2**	**No. of Families with Childrn in Parish School**
Did Not Contribute			861	$0	$0	$165,884,193	14
$ —	to	$ 100.00	431	$50	$21,550	$65,197,799	43
$ 100.00	to	$ 200.00	227	$150	$34,050	$18,950,834	61
$ 200.00	to	$ 300.00	218	$250	$54,500	$7,781,882	58
$ 300.00	to	$ 400.00	186	$350	$65,100	$1,471,179	54

$ 400.00	to	$ 500.00	145	$450	$65,250	$17,751	41
$ 500.00	to	$ 600.00	122	$550	$67,100	$1,504,903	41
$ 600.00	to	$ 700.00	90	$650	$58,500	$4,009,332	28
$ 700.00	to	$ 800.00	72	$750	$54,000	$6,966,791	27
$ 800.00	to	$ 900.00	62	$850	$52,700	$10,476,378	26
$ 900.00	to	$ 1,000.00	57	$950	$54,150	$14,887,642	20
$ 1,000.00	to	$ 1,500.00	191	$1,250	$238,750	$125,644,625	63
$ 1,500.00	to	$ 2,000.00	83	$1,750	$145,250	$142,667,832	21
$ 2,000.00	to	$ 2,500.00	45	$2,250	$101,250	$147,597,922	16
$ 2,500.00	to	$ 3,000.00	20	$2,750	$55,000	$106,820,362	5
$ 3,000.00	to	$ 3,500.00	13	$3,250	$42,250	$102,727,071	4
$ 3,500.00	to	$ 4,000.00	6	$3,750	$22,500	$65,778,880	2
$ 4,000.00	to	$ 4,500.00	5	$4,250	$21,250	$72,621,055	3
$ 4,500.00	to	$ 5,000.00	4	$4,750	$19,000	$74,341,101	2
$ 5,000.00	to	$ 10,000.00	7	$7,500	$52,500	$349,010,401	1
$ 10,000.00	to	$ 15,000.00	2	$12,500	$25,000	$290,938,543	0
		total	2847		$1,249,650	$1,775,296,475	
		Estimated mean	438.94			$623,567	
					Est std dev	$790	

16. A marketing study of 800 adults in the 18–34 age group reported the following information:
• Spent less than $100 on children's clothing per year: 75 responses
• Spent $100–$499 on children's clothing per year: 200 responses
• Spent $500–$999 on children's clothing per year: 25 responses
• The remainder reported spending nothing.
Estimate the sample mean and sample standard deviation of spending on children's clothing for this age group using formulas (2A.9) and (2A.11).

Answer:

Occurrences	Relative Frequency	Midpoint	Mid*Freq	Freq*(mid-Mean)^2
500	62.5%	0	0	6646.73
75	9.4%	50	3750	264.59
200	25.0%	300	60000	9689.94
25	3.1%	750	18750	13076.48
800	100.0%		82500	29677.73

Estimated mean: 82500/800 = $103.13

Estimated sample variance: 29677.73/(800-1) = $37.14

17. Data from the 2000 U.S. Census in the Excel file *California Census Data* show the distribution of ages for residents of California. Estimate the mean age and standard deviation of age for California residents using formulas (2A.9) and (2A.11), assuming these data represent a sample of current residents.

Answer:

Age	**Freq**	**Percent**	midpoint	mid*freq	Freq*(mid-mean)^2
Under 5 years	2,486,981	7.3%	2.50	6217452.5	2514617594
5 to 9 years	2,725,880	8.0%	7.00	19081160	2031272755
10 to 14 years	2,570,822	7.6%	12.00	30849864	1278214230
15 to 19 years	2,450,888	7.2%	17.00	41665096	733356282.1
20 to 24 years	2,381,288	7.0%	22.00	52388336	360147638.3
25 to 34 years	5,229,062	15.4%	29.50	154257329	120376977.2
35 to 44 years	5,485,341	16.2%	39.50	216670969.5	148438004.1
45 to 54 years	4,331,635	12.8%	49.50	214415932.5	1001044946
55 to 59 years	1,467,252	4.3%	57.00	83633364	756193826.4
60 to 64 years	1,146,841	3.4%	62.00	71104142	880087002
65 to 74 years	1,887,823	5.6%	69.50	131203698.5	2339354643
75 to 84 years	1,282,178	3.8%	79.50	101933151	2619773305
85 years and over	425,657	1.3%	90.00	38309130	1320691645
	33871648			1161729625	16103568849
			Est mean	34.30	475.429165
			Est std dev	21.80	

18. A deep-foundation engineering contractor has bid on a foundation system for a new world headquarters building for a Fortune 500 company. A part of the project consists of installing 311 auger cast piles. The contractor was given bid information for cost-estimating purposes, which consisted of the estimated depth of each pile; however, actual drill footage of each pile could not be determined exactly until construction was performed. The Excel file *Pile Foundation* contains the estimates and actual pile lengths after the project was completed. Compute the correlation coefficient between the estimated and actual pile lengths. What does this tell you?

Answer:

	Estimated pile length (ft)
Actual pile length (ft)	0.797

There is a high correlation between estimated and actual pile length. This means that the estimates made about the length were very accurate.

19. Call centers have high turnover rates because of the stressful environment. The national average is approximately 50%. The director of human resources for a large bank has compiled data from about 70 former employees at one of the bank's call centers (see the Excel file *Call Center Data*). For this sample, how strongly is length of service correlated with starting age?

Answer:

Length of service (years)

Starting age: -0.61

There is a moderately high negative correlation. This means that the younger workers are more likely to stay on the call center job longer.

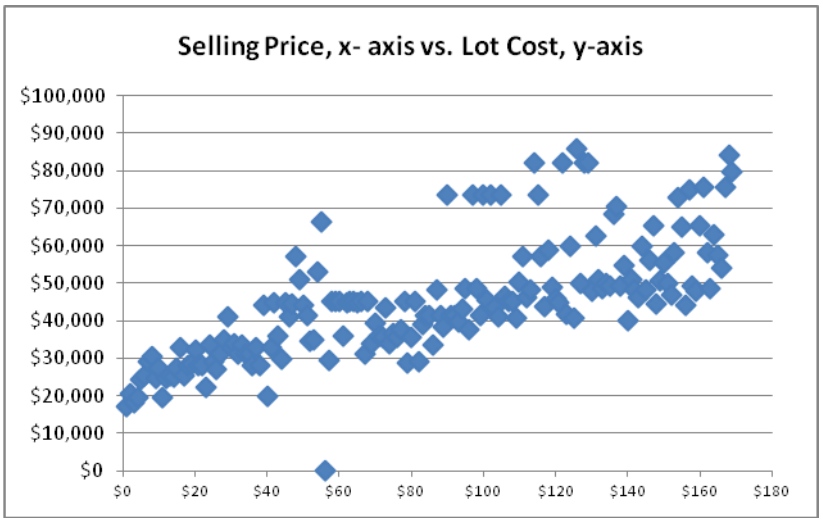
20. A national homebuilder builds single-family homes and condominium-style townhouses. The Excel file House Sales provides information on the selling price, lot cost, type of home, and region of the country (M = Midwest, S = South) for closings during one month.

a. Construct a scatter diagram showing the relationship between sales price and lot cost. Does there appear to be a linear relationship? Compute the correlation coefficient.

b. Construct scatter diagrams showing the relationship between sales price and lot cost for each region. Do linear relationships appear to exist? Compute the correlation coefficients.

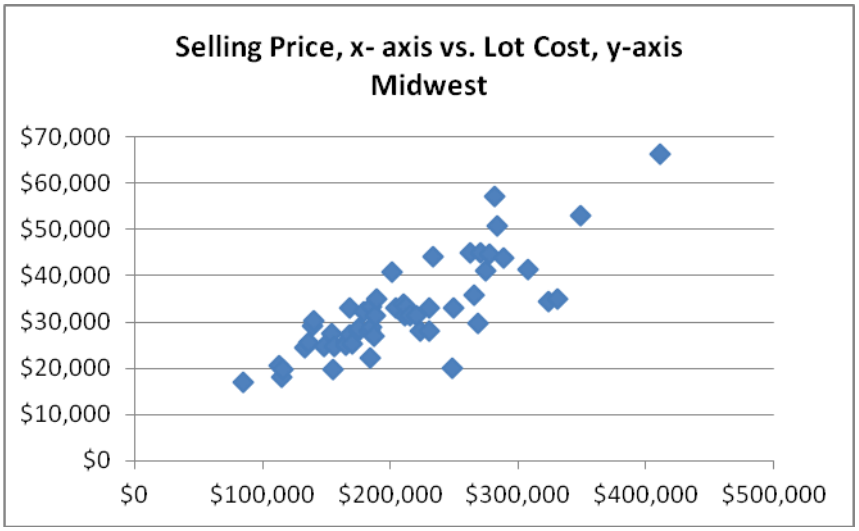
c. Construct scatter diagrams showing the relationship between sales price and lot cost for each type of house. Do linear relationships appear to exist? Compute the correlation coefficients.

Answer:

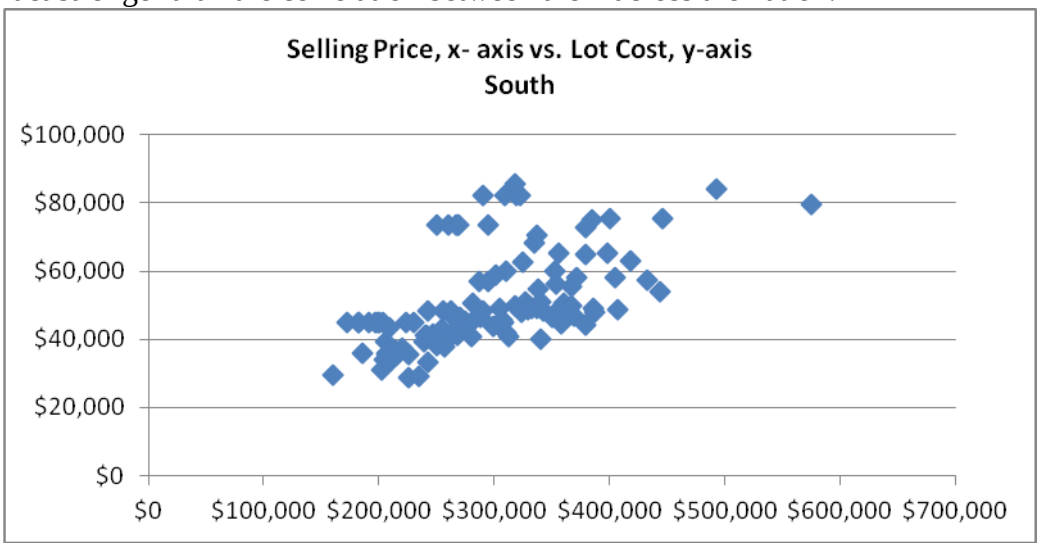


Correlation between Lot cost and Selling price is 0.73
There is a strong linear relationship between lot cost and selling price.

b.

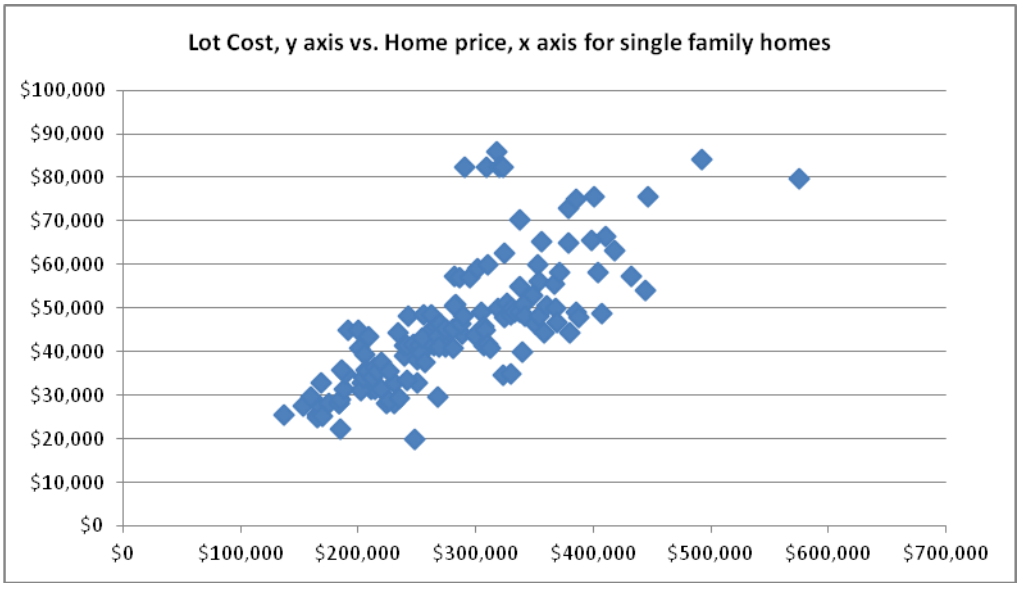


Midwest
Correlation between Lot cost and Selling price is 0.81
There is a strong linear relationship between lot cost and selling price within Midwest, in fact stronger than the correlation between the 2 across the nation.

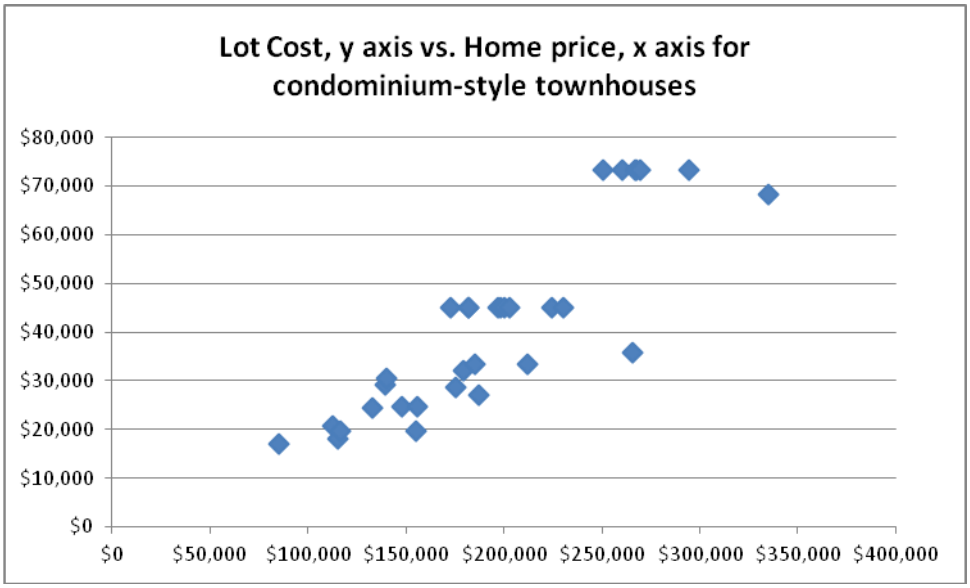


South
Correlation between Lot cost and Selling price is 0.57
There is a moderate linear relationship between lot cost and selling price, in the south, and weaker than relationship in the midwest.

c.



Single family homes
Correlation between Lot cost and Selling price is 0.751
There is a strong linear relationship between lot size and house price for single family homes.



Condominium-style townhouses
Correlation between Lot cost and Selling price is 0.877
There is a strong linear relationship between lot size and house price for condominium-style townhouses.

21. The Excel file *Infant Mortality* provides data on infant mortality rate (deaths per 1,000 births), female literacy (percentage who read), and population density (people per square kilometer) for 85 countries. Compute the correlation matrix for these three variables. What conclusions can you draw?

Answer:

	Mortality	**Density**	**Literacy**
Mortality	**1**		
Density	**-0.1853097**	**1**	
Literacy	**-0.8434115**	**0.0286499**	**1**

Infant mortality has extremely high negative correlation with educational level, -.85, which indicates that the lower the educational level of the mother, the higher the infant mortality rate. Mortality has only a slight negative correlation with population density, which indicates that higher population density, such as in urban areas, has a slightly lower rate of infant mortality.

22. The Excel file *Refrigerators* provides data on various brands and models. Compute the correlation matrix for the variables. What conclusions can you draw?

Answer:

	Price	**Storage Capacity (cu.ft.)**	**Energy consumption KWH/YR.**	**Energy efficiency (kWh/YR/cu.ft.)**
Price	1			
Storage Capacity (cu.ft.)	-0.04067108	1		
Energy consumption KWH/YR.	0.123752728	0.130492159	1	
Energy efficiency (kWh/YR/cu.ft.)	0.123684737	-0.409933731	0.847473415	1

Price is not related to storage capacity, and refrigerators with higher energy consumption and efficiency are only slightly more expensive. Refrigerators with higher storage capacity tend to consume slightly more energy, but they are definitely much less efficient, with a negative correlation of -.41. Energy consumption has a high positive correlation with energy efficiency, .85, the strongest correlation in the comparison. The refrigerators with the highest energy consumption are also the most efficient.

23. The worksheet Mower Test in the Excel file Quality Measurements shows the results of testing 30 samples of 100 lawn mowers prior to shipping. Find the proportion of units that failed the test for each sample. What proportion failed overall?

Answer:

Sample	Proportion
1	3%
2	4%
3	1%
4	0%
5	1%
6	5%
7	2%
8	1%
9	0%
10	2%
11	2%
12	3%
13	3%
14	1%
15	1%
16	2%
17	2%
18	3%
19	2%
20	4%
21	2%
22	1%
23	1%
24	2%
25	1%
26	0%
27	2%
28	1%
29	0%
30	2%
Overall	1.80%

24. The Excel file *EEO Employment Report* shows the number of people employed in different professions for various racial and ethnic groups. Find the proportion of men and women in each ethnic group for the total employment and in each profession.

Answer:

Racial/Ethnic Group and Gender	Total Employment	Officials & Managers	Professionals	Technicians	Sales Workers	Office & Clerical Workers	Craft Workers	Operatives	Laborers	Service Workers
Men - All	55%	69%	49%	50%	38%	17%	91%	70%	65%	39%
Women - All	45%	31%	51%	50%	62%	83%	9%	30%	35%	61%
Men - White	58%	71%	52%	56%	43%	17%	93%	75%	67%	40%
Women - White	42%	29%	48%	44%	57%	83%	7%	25%	33%	60%
Men - Minority	50%	58%	38%	35%	29%	16%	84%	64%	64%	37%
Women - Minority	50%	42%	62%	65%	71%	84%	16%	36%	36%	63%
Men - Black	46%	55%	29%	32%	28%	16%	82%	63%	61%	35%
Women - Black	54%	45%	71%	68%	72%	84%	18%	37%	39%	65%
Men - Hispanic	71%	75%	67%	69%	42%	26%	93%	77%	69%	69%
Women - Hispanic	29%	25%	33%	31%	58%	74%	7%	23%	31%	31%
Men - Asian American	56%	73%	64%	63%	36%	23%	85%	50%	46%	37%
Women - Asian American	44%	27%	36%	37%	64%	77%	15%	50%	54%	63%
Men - American Indian	62%	73%	58%	63%	35%	20%	89%	70%	76%	40%
Women - American Indian	38%	27%	42%	37%	65%	80%	11%	30%	24%	60%

25. A mental health agency measured the self-esteem score for randomly selected individuals with disabilities who were involved in some work activity within the past year. The Excel file *Self Esteem* provides the data, including the individuals' marital status, length of work, type of support received (direct support includes job-related services such as job coaching and counseling), education, and age. Construct a cross-tabulation of the number of individuals within each classification of marital status and support level.

Answer:

# of Individuals	Support Level		
Marital Status	Direct	None	Total
Divorced	2	12	14
Married	5	3	8
Separated	7	8	15
Single	16	7	23
Total	30	30	60

26. Construct cross-tabulations of Gender versus Carrier and Type versus Usage in the Excel file Cell Phone Survey. What might you conclude from this analysis?

Answer:

# of Individuals	**Carrier**					
Gender	AT&T	Sprint	Verizon	T-mobile	Other	**Total**
Female	8	2	5	1	2	18
Male	18	3	5	1	7	34
Total	26	5	10	2	9	52

# of Individuals	**Usage**				
Type	Very high	High	Average	Low	**Total**
Basic	2	0	6	4	12
Camera	7	0	11	1	19
Smart	13	2	6	0	21
Total	22	2	23	5	52

Both males and females prefer AT&T than any other carriers.
Smart types are the highest.

27. The Excel file Unions and Labor Law Data reports the percentage of public and private sector employees in unions in 1982 for each state, along with indicators of whether the states had a bargaining law that covered public employees or right - to - work laws.

a. Compute the proportion of employees in unions in each of the four categories: public sector with bargaining laws, public sector without bargaining laws, private sector with bargaining laws, and private sector without bargaining laws.

b. Compute the proportion of employees in unions in each of the four categories: public sector with right to - work laws, public sector without right - to - work laws, private sector with right - to - work laws, and private sector without right - to - work laws.

c. Construct a cross - tabulation of the number of states within each classification of having or not having bargaining laws and right - to - work laws.

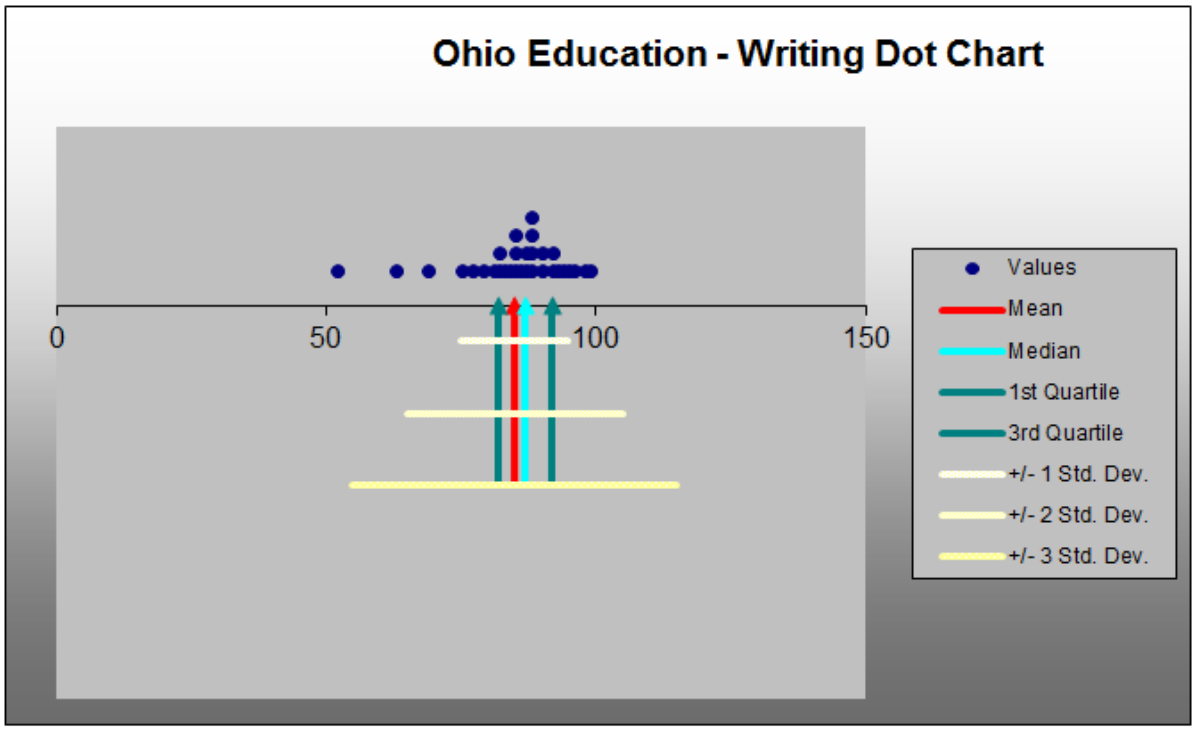
Answer:

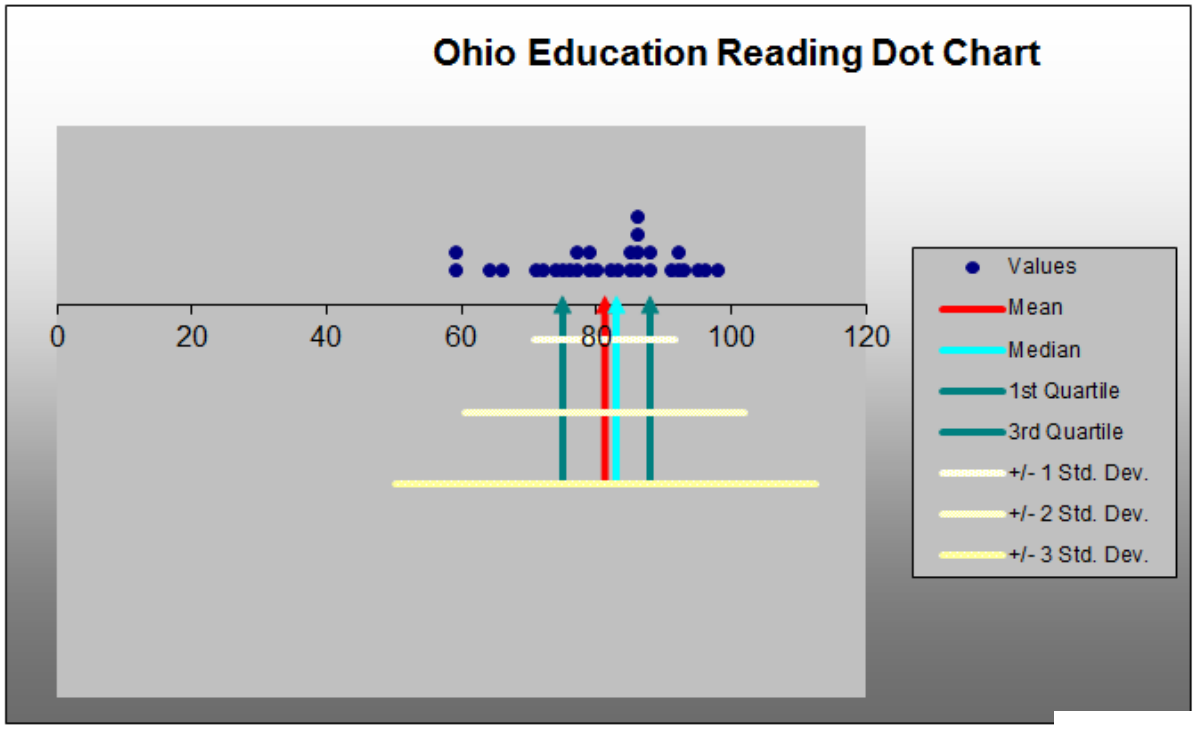
Right to work laws	**Bargaining Laws**	
Do Not Exist	Do not exist	Do exist
Average % of Public sector employees in Unions	28.20	43.20
Average % of Private sector employees in Unions	18.75	20.34
Do Exist		
Average % of Public sector employees in Unions	27.25	23.70
Average % of Private sector employees in Unions	11.00	9.33

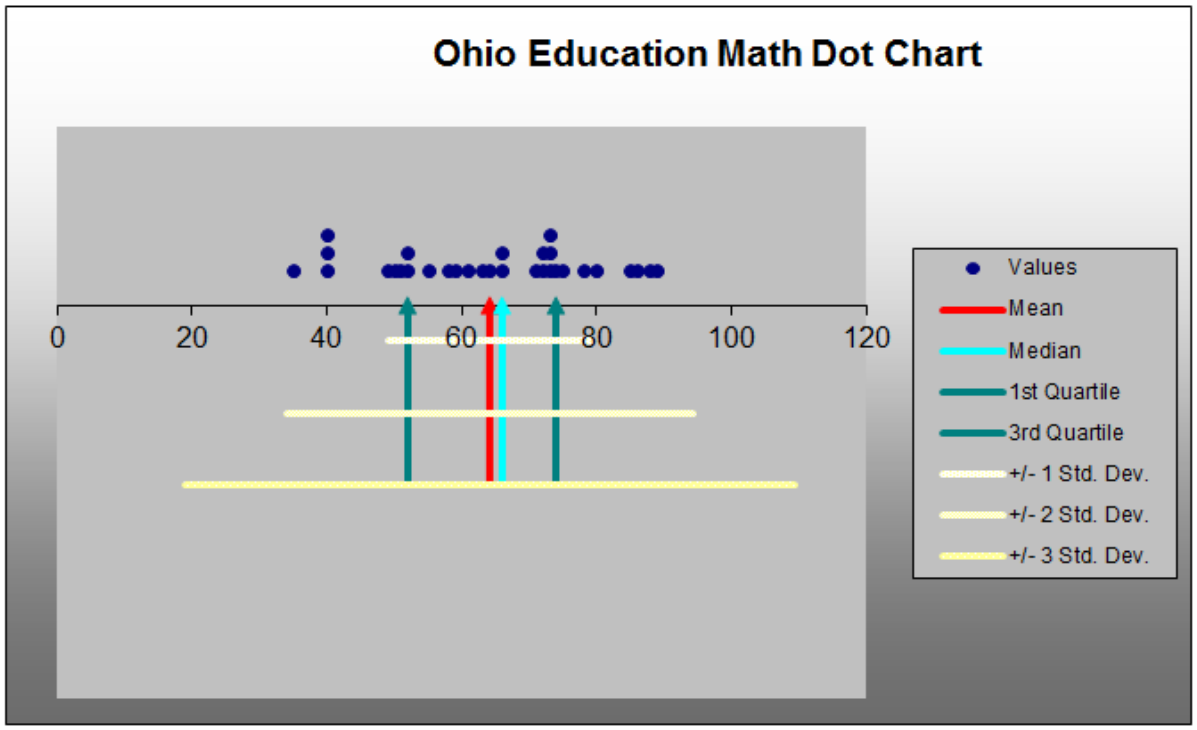
# of States	**Bargaining Laws**		
Right to work laws	Do not exist	Do exist	Total
Do Not Exist	10	20	30
Do Exist	13	7	20
Total	23	27	50

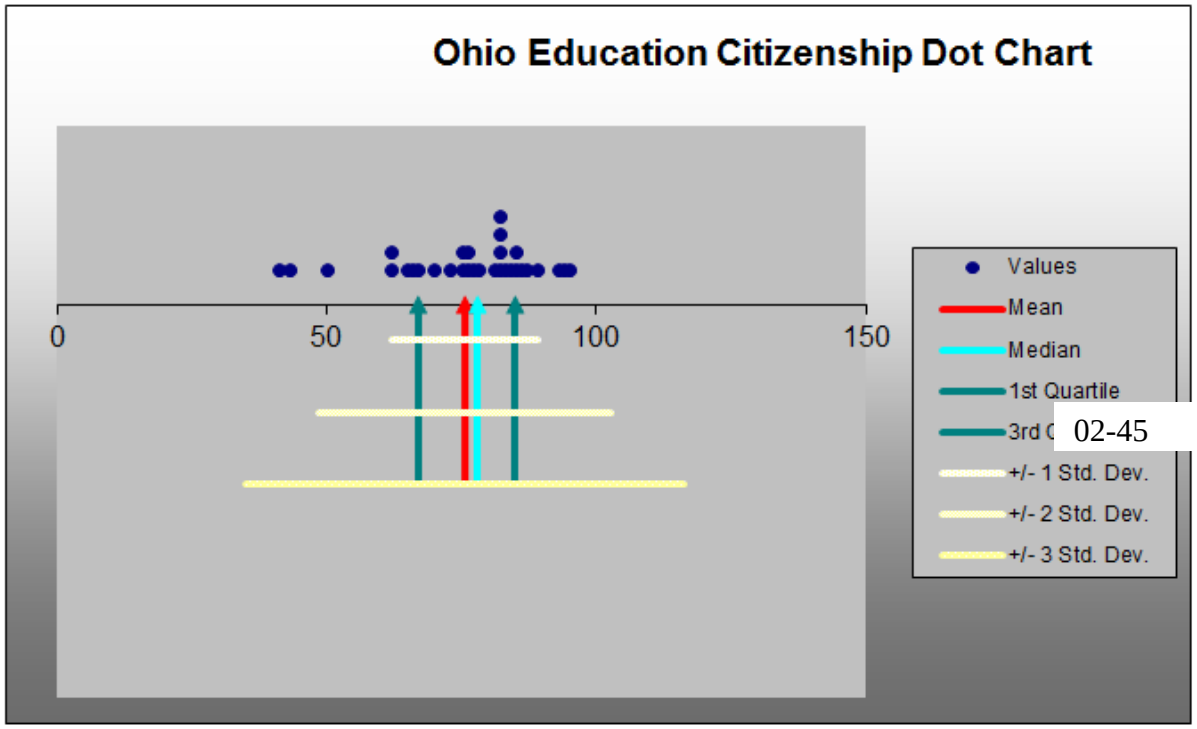
28. Construct box plots and dot-scale diagrams for each of the variables in the data set Ohio Education Performance. What conclusions can you draw from them? Are any possible outliers evident?

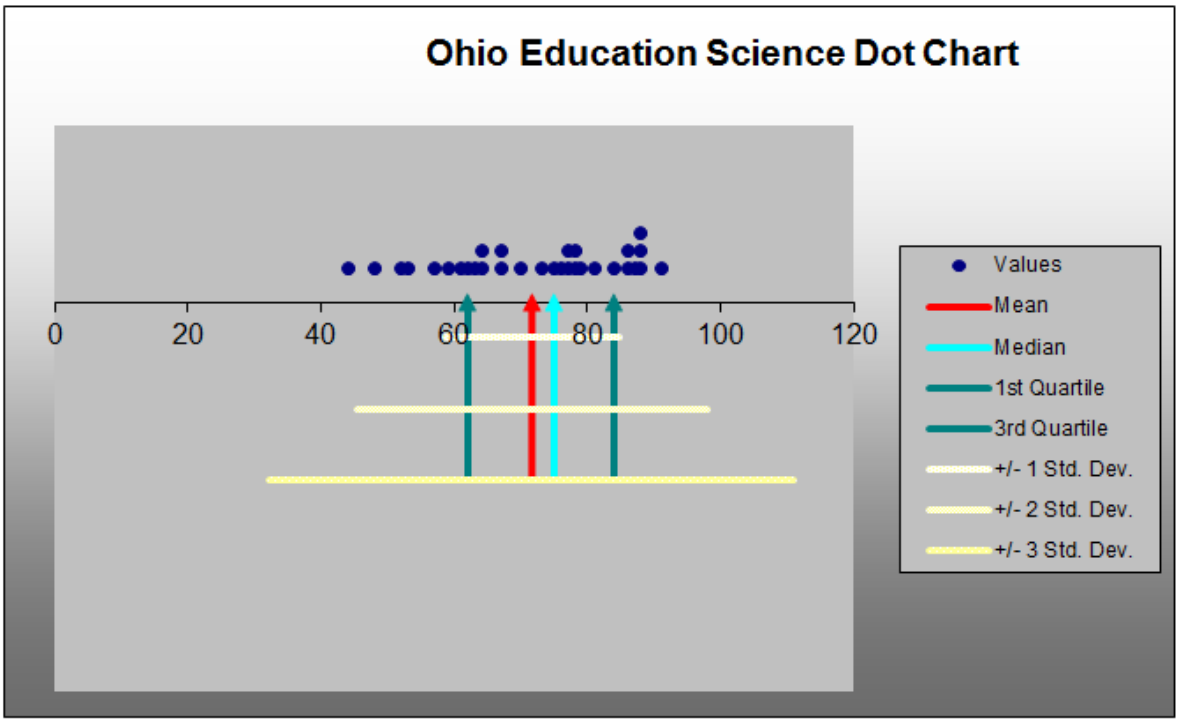
Answer:

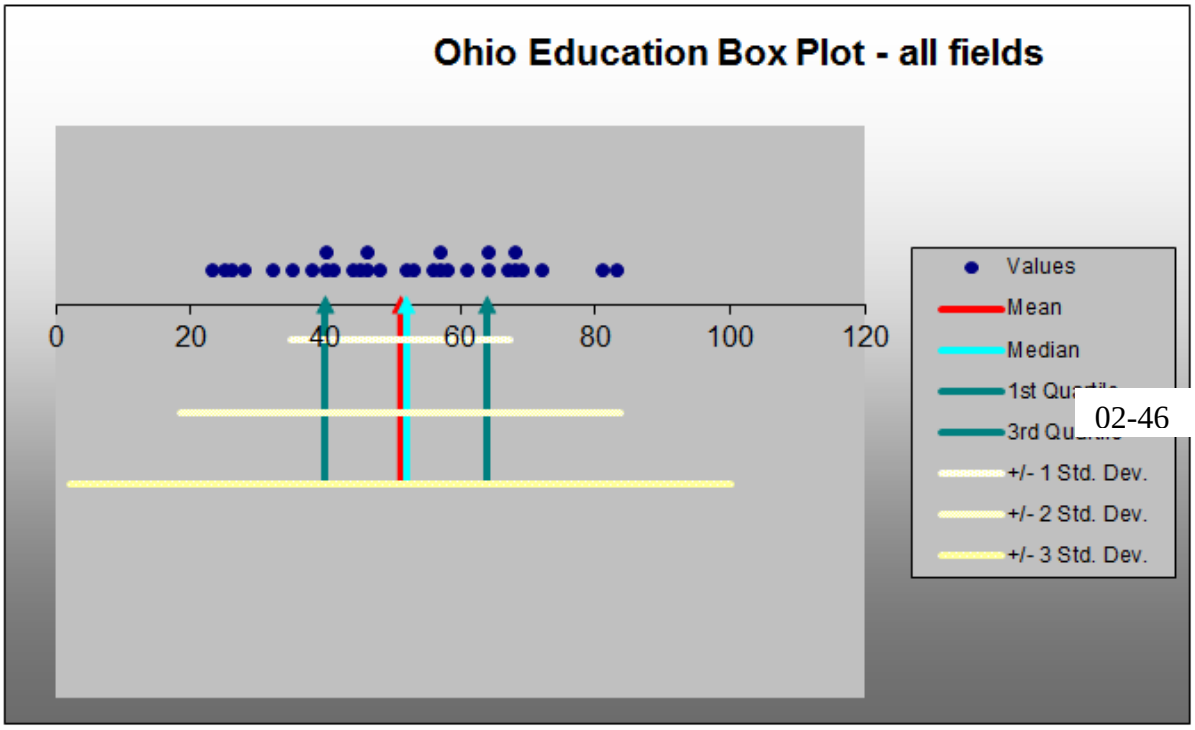


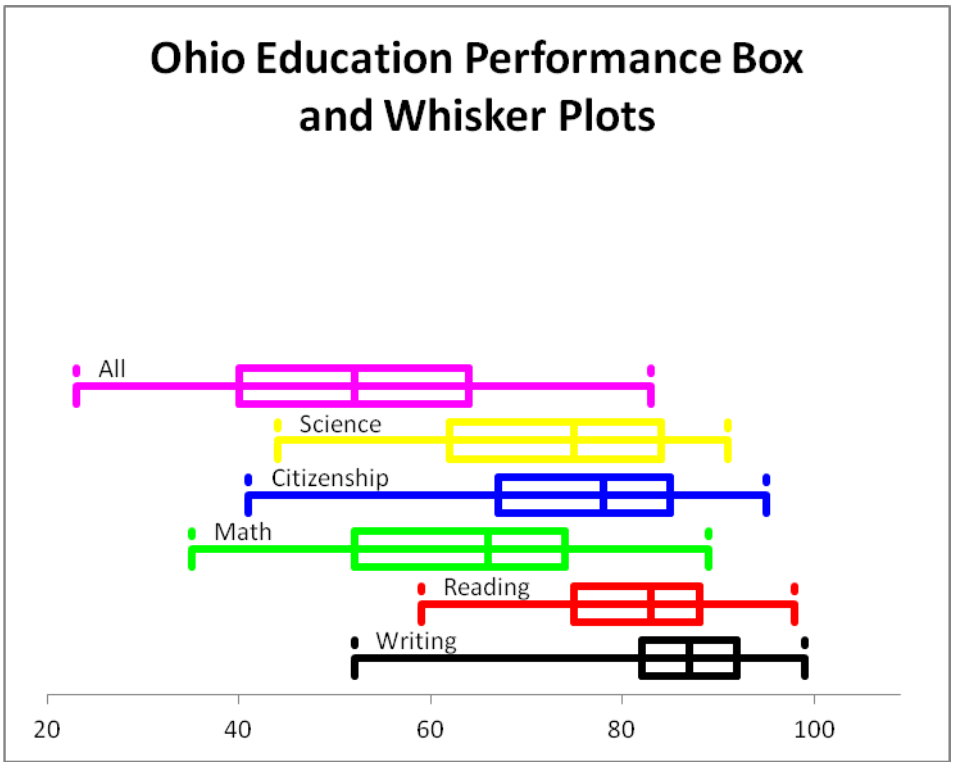












02-47

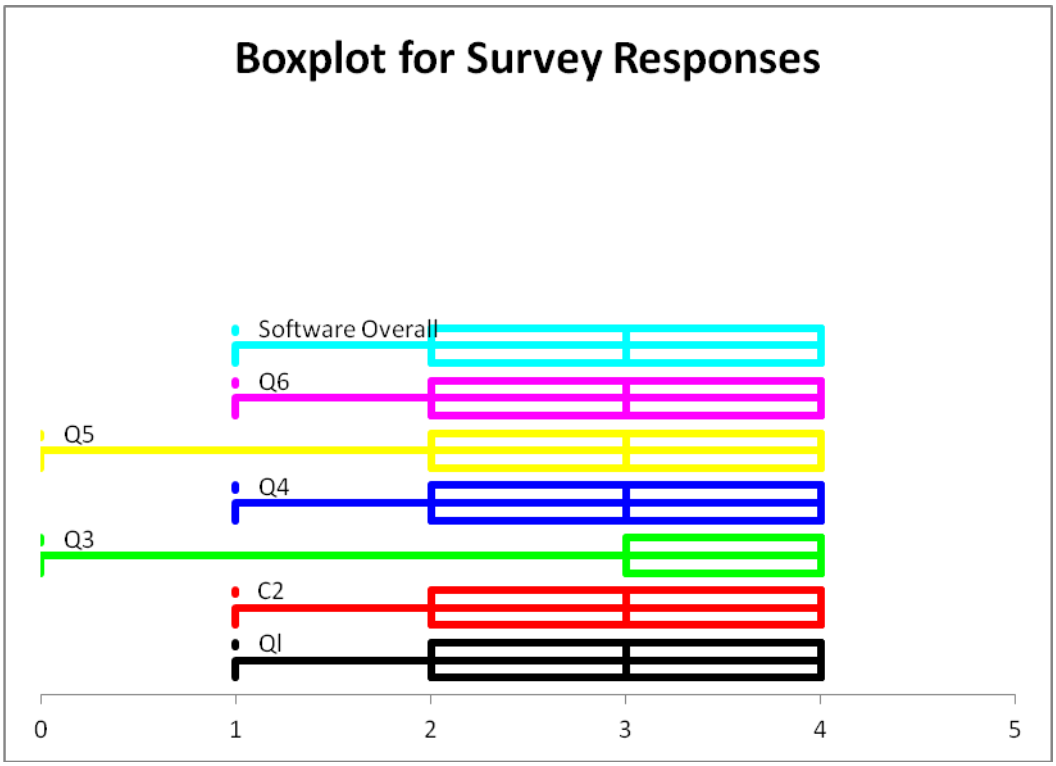
Each of the individual field categories are skewed to the left. The overall category appears to not be skewed. Reading and writing scores are the highest, followed by citizenship and science. Math scores are the lowest.

On the dot charts, the math scores are very spread out and the writing scores are concentrated together.

29. A producer of computer‐aided design software for the aerospace industry receives numerous calls for technical support. Tracking software is used to monitor response and resolution times. In addition, the company surveys customers who request support using the following scale:
0—Did not exceed expectations
1—Marginally met expectations
2—Met expectations
3—Exceeded expectations
4—Greatly exceeded expectations
The questions are as follows:
Q1: Did the support representative explain the process for resolving your problem?
Q2: Did the support representative keep you informed about the status of progress in resolving your problem?
Q3: Was the support representative courteous and professional?
Q4: Was your problem resolved?
Q5: Was your problem resolved in an acceptable amount of time?
Q6: Overall, how did you find the service provided by our technical support department?
A final question asks the customer to rate the overall quality of the product using this scale:
0—Very poor
1—Poor
2—Good
3—Very good
4—Excellent

A sample of survey responses and associated resolution and response data are provided in the Excel file Customer Support Survey. Use descriptive statistics, box plots, and dot‐scale diagrams as you deem appropriate to convey the information in these sample data and write a report to the manager explaining your findings and conclusions.

Answer:



Stem and Leaf Diagram for Response Time

Stem unit: **10**

0	**2 2 2 2 5 6 8**
1	**1 1 1 2 2 2 3 3 3 3 5 5 5 7 9 9 9**
2	**0 0 1 1 2 4 5 5 5 7 8 9**
3	**1 3 3 9**
4	**7 8**
5	
6	**1**
7	
8	
9	
10	
11	**8**

Statistics	
Sample Size	44
Mean	21.9091
Median	19
Std. Deviation	19.4862
Minimum	2
Maximum	118

All of the questions on the box and whisker plot indicate that the customers' expectations were generally met or exceeded because they are all skewed to the left.
Questions 3 and 5 had some dissatisfied customers, indicating that there is a need for improvement in the interpersonal skills of the support representatives and the amount of time to resolve a problem.

The answers to the overall satisfaction question about the software indicate that the customers are satisfied.
The dot scale chart for problem resolution time indicates that the distribution is strongly skewed to the right with half of the problems resolved within the 10 -1 5 minute time range. 75% of the problems were resolved within 5 days, and very few took several weeks or longer.
The stem and leaf plot show that half of the response times were within 20 minutes of the call. Most of the responses are within 10-20 minutes, with a large number in the 20-30 minute category. One call took 1 hour and another took 2 hours. The causes for this excessive amount of time should be investigated.

30. Call centers have high turnover rates because of the stressful environment. The national average is approximately 50%. The director of human resources for a large bank has compiled data from about 70 former employees at one of the bank's call centers (see the Excel file *Call Center Data*). Use PivotTables to find these items:
a. The average length of service for males and females in the sample.
b. The average length of service for individuals with and without a college degree.
c. The average length of service for males and females with and without prior call center experience.
What conclusions might you reach from this information?

Answer:

Gender	Average length of service (years)		
Female	2.01		
Male	1.76		
Overall total	1.89		
College degree?	Average length of service (years)		
No	2.06		
Yes	1.52		
Overall total	1.89		
	Prior Call Center Experience		
Average length of service (years)	No	Yes	Overall total
Female	2.18	1.77	2.01
Male	1.78	1.72	1.76
Overall total	1.99	1.75	1.89

Female employees tend to work about 3 months longer.
Employees without a college degree work about 6 months longer.
Employees with prior call center experience tend to serve about 3 months less.
Out of all employees with prior call center experience, the length of service for males and females is very similar.
For employees without call center experience, female employees tend to serve 3 - 4 months longer.

31. The Excel file *University Grant Proposals* provides data on the dollar amount of proposals, gender of the researcher, and whether the proposal was funded or not. Construct a PivotTable to find the average amount of proposals by gender and outcome.

Answer:

	Female	Male	Overall total
Funded	$473415.19	$351441.97	$378563.07
Rejected	$713935.60	$918235.60	$866250.35
Overall average	$612011.03	$653277.62	$643382.51

32. A national homebuilder builds single - family homes and condominium - style townhouses. The Excel file House Sales provides information on the selling price, lot cost, type of home, and region of the country (M = Midwest, S = South) for closings during one month. Use PivotTables to find the average selling price and lot cost for each type of home in each region of the market. What conclusions might you reach from this information?

Answer:

Row Labels	**Sum of Selling Price**	**Average Selling Price**
Midwest	**11532830**	$209,687.82
Single Family	9026998	$231,461.49
Townhouse	2505832	$156,614.50
South	**33378706**	$295,386.78
Single Family	29644637	$305,614.81
Townhouse	3734069	$233,379.31
Grand Total	**44911536**	$267,330.57

In the Midwest, single family homes are almost 50% more expensive, and in the South, single family homes are only about30% more expensive.

Single family homes in the south are about $75000 more expensive than in the midwest.

Row Labels	**Sum of Lot Cost**	**Average lot price**
Single Family	**6240047**	$45,882.70
Midwest	1366239	$35,031.77
South	4873808	$50,245.44
Townhouse	**1332670**	$41,645.94
Midwest	418670	$26,166.88
South	914000	$57,125.00
Grand Total	**7572717**	$45,075.70

Average lot prices are more expensive in the south, as well as the total home prices are more in the south.

33. The Excel file *MBA Student Survey* provides data on a sample of students' social and study habits. Use PivotTables to find the average age, number of nights out per week, and study hours per week by gender, whether the student is international or not, and undergraduate concentration.

Answer:

Gender	Age	Nights out/week	Study hours/week
Female	37	5	10
Female	23	2	50
Female	22	3	10
Female	32	1	20
Female	22	1	6
Female	26	0.5	15
Female	26	1	14
Female	23	2	22
Female	24	3	14
Female	28	2	20
Female	25	1	15
Female	40	1	28
Female	22	4	10
Female	27	2	10
Female	28	1	28
Male	28	2	6
Male	24	2	20
Male	32	2	25
Male	23	2	15
Male	22	3	12
Male	32	1	14
Male	24	2	20
Male	22	2	10
Male	24	1.5	3
Male	31	1	25
Male	25	1	16
Male	26	0	15
Male	24	2	20
Male	28	1	30
Male	26	7	60
Male	37	0	7
Male	23	1	10
Male	24	3	18
Male	23	1	15
Male	26	3	15

Row Labels	Sum of Age	Sum of Nights out/week	Sum of Study hours/week

Female	405	29.5	272
Male	524	37.5	356
Grand Total	**929**	**67**	**628**

Row Labels	Average Age	Average Nights out/week	Average Study hours/week
Female	27	1.97	18.13
Male	26.2	1.88	17.8
Grand Total	26.54	1.91	17.94

Men and women go out about 2 nights/week and study about 18 hours per week.

International?

Female	15	Total
No	7	The female mix is about 50% each between international and domestic.
Yes	8	
Male	20	Total
No	15	For males, about 75% of the men are domestic vs. international in the MBA program.
Yes	5	
Grand Total	35	

Row Labels

Female	**15**
Business	4
Engineering	3
Liberal Arts	1
Other	3
Sciences	4

Male	20
Business	5
Engineering	6
Liberal Arts	5
Other	3
Sciences	1
Grand Total	35

Total 18 out of the 35 students were either from business or engineering backgrounds. The fewest MBA's had science backgrounds. Gender mix was even, except more women came from science.

34. A mental health agency measured the self‐esteem score for randomly selected individuals with disabilities who were involved in some work activity within the past year. The Excel file Self‐Esteem provides the data including the individuals' marital status, length of work, type of support received (direct support includes job‐related services such as job coaching and counseling), education, and age. Use PivotTables to find the average length of work and self‐esteem score for individuals in each classification of marital status and support level. What conclusions might you reach from this information?

Answer:

Row Labels	**Sum of Length of Work (months)**	**Sum of Self Esteem**
Divorced	**121**	**47**
Direct	41	8
None	80	39
Married	**164**	**34**
Direct	133	24
None	31	10
Separated	**234**	**54**
Direct	155	27
None	79	27
Single	**457**	**92**
Direct	405	65
None	52	22
Grand Total	**976**	**222**

Row Labels	**Average Length of Work (months)**	**Average Self Esteem**
Divorced	8.64	3.4
Direct	20.50	4
None	6.67	3.3
Married	20.5	4.3
Direct	26.6	4.8
None	10.33	3.3
Separated	15.6	3.6
Direct	22.14	3.9
None	9.88	3.4
Single	19.87	4
Direct	25.31	4.4
None	7.43	3.1
Grand Total	16.27	3.76

Those with direct support had higher self-esteem in every category of marital status. They also worked about twice as long in every marital status category if they had direct support. Married people worked the longest and had the highest average self-esteem.

35. The Excel file *Cell Phone Survey* reports opinions of a sample of consumers regarding the signal strength, value for the dollar, and customer service for their cell phone carriers. Use PivotTables to find the following:

a. The average signal strength by type of carrier.

b. Average value for the dollar by type of carrier and usage level.

c. Variance of perception of customer service by carrier and gender.

What conclusions might you reach from this information?

Answer:

Row Labels	**Sum of Signal strength**	Average signal strength
AT&T	90	3.46
Other	24	2.67
Sprint	14	2.8
T-mobile	6	3
Verizon	38	3.8
Grand Total	**172**	3.31

The signal strength for Verizon and AT&T is the strongest. Sprint and other services have the weakest signal strengths.

Row Labels	**Sum of Value for the Dollar**	**Average Value for the Dollar**
Average	**78**	**3.39**
AT&T	35	3.50
Other	17	2.83
Sprint	14	4.67
T-mobile	4	4
Verizon	8	2.67
High	**8**	**4**
AT&T	8	4
Low	**16**	**3.2**
AT&T	5	2.5
Other	4	4
T-mobile	4	4
Verizon	3	3
Very high	**76**	**3.45**
AT&T	36	3
Other	6	3
Sprint	10	5
Verizon	24	4
Grand Total	**178**	**3.42**

Sprint, T-mobile and other services had the best value for the dollar.
Verizon and AT&T had the lowest value for the dollar.

AT&T had the most high and very high users of the service. Verizon had also many very high users, about 1/3.

Variance of perception of customer service by carrier

Row Labels	Var of Customer Service
AT&T	0.99
Other	0.78
Sprint	0.2
T-mobile	0.5
Verizon	1.33
Grand Total	**0.93**

Verizon has the highest variance of perception of customer service. Sprint has the lowest variance of perception of customer service.

Variance of perception of customer service by gender

Row Labels	Var of Customer Service
F	0.76
M	0.97
Grand Total	**0.93**

Variance of perception of customer service is higher in males than females

36. The Excel file Freshman College Data shows data for four years at a large urban university. Use PivotTables to examine differences in student high school performance and first - year retention among different colleges at this university. What conclusions do you reach?

Answer:

Row Labels	Sum of HS GPA	Sum of 1st year retention rate
Architecture	14.42855	3.556529915
Business	12.97816	2.970274854
Education	12.38369	2.805754386
Engineering	14.21823	3.271706161
Health Sciences	12.821214	2.729490196
Liberal Arts	12.614	2.755470588
Music	13.6531	3.29533121
North Central Branch Campus	10.4392299	2.538784195
Nursing	13.35616	3.024490196
South Central Branch Campus	10.848224	2.521784195
Vocational Technology	11.12681	2.740983425

Grand Total	138.8673679	32.21059932

Row Labels	Average HS GPA	Average 1st year retention rate
Architecture	3.61	89%
Business	3.24	74%
Education	3.10	70%
Engineering	3.55	82%
Health Sciences	3.21	68%
Liberal Arts	3.15	69%
Music	3.41	82%
North Central Branch Campus	2.61	63%
Nursing	3.34	76%
South Central Branch Campus	2.71	63%
Vocational Technology	2.78	69%
Grand Total	3.16	73%

FIGURE: Baldrige Examination Scores

	A	B	C	D	E	F	G	H	I	J	K
1	Baldrige Examination Scores										
2											
3			Individual Assessment Percentage Scores								
4	Item	Maximum	Examiner	Examiner	Examiner	Examiner	Examiner	Examiner	Examiner	Examiner	Consensus
5		Points	1	2	3	4	5	6	7	8	Score
6	1.1	70	80	80	50	60	60	70	70	50	75
7	1.2	50	30	50	30	40	40	60	60	50	50
8	2.1	40	50	70	50	50	40	60	70	40	65
9	2.2	45	30	40	50	50	60	40	30	50	55
10	3.1	40	30	60	40	60	50	30	50	30	45
11	3.2	45	30	50	60	60	60	50	30	60	50
12	4.1	45	40	70	50	60	40	30	20	50	50
13	4.2	45	30	20	40	40	30	30	10	30	40
14	5.1	45	70	50	60	40	40	60	60	50	60
15	5.2	40	50	20	40	40	70	40	40	20	40
16	6.1	35	50	60	50	50	50	40	30	40	45
17	6.2	50	40	40	60	50	40	30	60	50	50
18	7.1	100	60	70	70	70	80	70	70	70	75
19	7.2	70	50	60	70	50	70	50	70	70	70
20	7.3	70	50	40	50	50	70	30	30	50	50
21	7.4	70	40	50	50	50	50	40	20	60	45
22	7.5	70	70	70	60	70	50	60	80	50	75
23	7.6	70	60	80	70	60	70	40	60	70	70
24	Weighted score		499.5	565.5	546.5	543	564	478.5	503	523	585

The criteria undergo periodic revision, so the items and maximum points will not necessarily coincide with the current year's criteria.

Item	Item weight	Consensus Score	Mean	Median	Range	StDev	CV	Skew	Consensus - Median	Group Weights
1.1	7%	75	65.0	65	30	12.0	0.18	0.00	10	58%
1.2	5%	50	45.0	45	30	12.0	0.27	0.00	5	42%
2.1	4%	65	53.8	50	30	11.9	0.22	0.39	15	47%
2.2	5%	55	43.8	45	30	10.6	0.24	-0.04	10	53%
3.1	4%	45	43.8	45	30	13.0	0.30	0.11	0	47%
3.2	5%	50	50.0	55	30	13.1	0.26	-1.02	-5	53%
4.1	5%	50	45.0	45	50	16.0	0.36	0.00	5	50%
4.2	5%	40	28.8	30	30	9.9	0.34	-0.86	10	50%
5.1	5%	60	53.8	55	30	10.6	0.20	-0.04	5	53%
5.2	4%	40	40.0	40	50	16.0	0.40	0.55	0	47%
6.1	4%	45	46.3	50	30	9.2	0.20	-0.49	-5	41%
6.2	5%	50	46.3	45	30	10.6	0.23	0.04	5	59%
7.1	10%	75	70.0	70	20	5.3	0.08	0.00	5	22%
7.2	7%	70	61.3	65	20	9.9	0.16	-0.31	5	16%
7.3	7%	50	46.3	50	40	13.0	0.28	0.41	0	16%
7.4	7%	45	45.0	50	40	12.0	0.27	-1.34	-5	16%
7.5	7%	75	63.8	65	30	10.6	0.17	-0.04	10	16%
7.6	7%	70	63.8	65	40	11.9	0.19	-0.97	5	16%
Sum	1,000									

The median is used as a typical examiner's score, since there are few items (3.2, 4.2, 7.6) with strongly left-skewed distributions, indicating that the median would be significantly higher than the mean. Item 7.1 has very small coefficient of variation, indicating general agreement on this aspect of the company's business results, and it also got the highest median score among all the items. Most of the other high median scoring items are in the same business results group. On the other hand, group 4 on measurement, analysis, and knowledge management has relatively high level of disagreement among the examiners. Items 4.1 in the measurement group and 5.2 in workforce focus groups have a pretty wide range of examiner's scores, perhaps indicating few dissenters.

Consensus Score	
Groups	**Mean**
1 - Leadership	64.6
2 - Planning	59.7
3 - Customer	47.6
4 - Analysis	45.0
5 - Workforce	50.6
6 - Process	47.9
7 - Results	64.9
Overall	58.5

Group analysis confirms the item analysis results, with business results and leadership scoring the highest, and analysis the lowest, closely followed by customer focus and process management as three potential areas of concern.

The colleges of business, architecture, engineering, music and nursing had the highest retention levels.

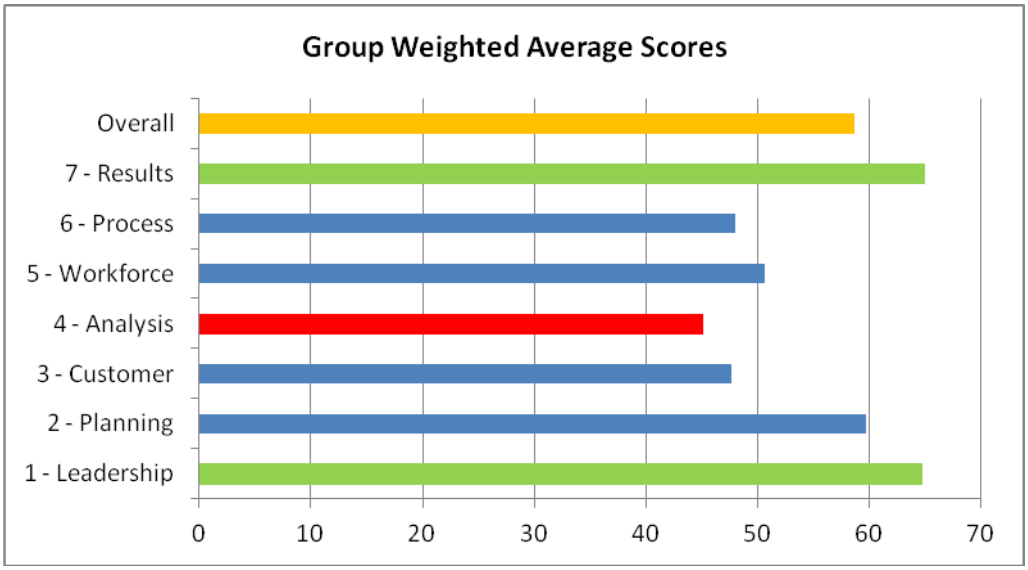
The colleges with the highest incoming high school grades were architecture, business, engineering, health sciences, music and nursing, almost all the same as those with the highest retention rates, except for health sciences.

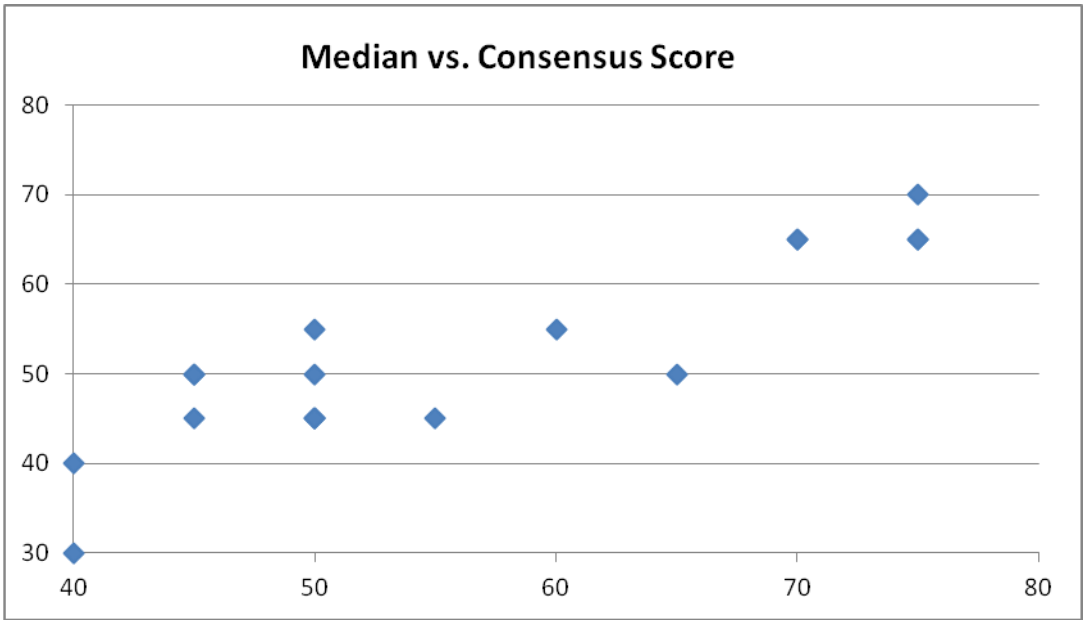
The Malcolm Baldrige Award recognizes U.S. companies that excel in high‐performance management practice and have achieved outstanding business results. The award is a public–private partnership, funded primarily through a private foundation and administered through the National Institute of Standards and Technology (NIST) in cooperation with the An 02-60 Society for Quality (ASQ) and is presented annually by the President of the United States. It was created to increase the awareness of American business for quality and good business practices and has become a worldwide standard for business excellence. See the Program Web site at www.nist.gov/baldrige for more information. The award examination is based on a rigorous set of criteria, called the Criteria for Performance Excellence, which consists of seven major categories: Leadership; Strategic Planning; Customer Focus; Measurement, Analysis, and Knowledge Management; Workforce Focus; Process Management; and Results. Each category consists of several items that focus on major requirements on which businesses should focus. For example, the two items in the Leadership category are Senior Leadership and Governance and Social Responsibilities. Each item, in turn, consists of a small number of areas to address, which seek specific information on approaches used to ensure and improve competitive performance, the deployment of these approaches, or results obtained from such deployment. The current year's criteria may be downloaded from the Web site. Applicants submit a 50‐page document that describes their management practices and business results that respond to the criteria. The evaluation of applicants for the award is conducted by a volunteer board of examiners selected by NIST. In the first stage, each application is reviewed by a team of examiners. They evaluate the applicant's response to each criteria item, listing major strengths and opportunities for improvement relative to the criteria. Based on these comments, a score from 0 to 100 in increments of 10 is given to each item.

Scores for each examination item are computed by multiplying the examiner's score by the maximum point value that can be earned for that item, which varies by item. These point values weight the importance of each item in the criteria. Then the examiners share information on a secure Web site and discuss issues via telephone conferencing to arrive at consensus comments and scores. The consensus stage is an extremely important step of the process. It is designed to smooth out variations in examiners' scores, which inevitably arise because of different perceptions of the applicants' responses relative to the criteria, and provide useful feedback to the applicants. In many cases, the insights of one or two judges may sway opinions, so consensus scores are not simple averages. A national panel of judges then reviews the scores and selects the highest‐scoring applicants for site visits.

At this point, a team of examiners visits the company for the greater part of a week to verify information contained in the written application and resolve issues that are unclear or about which the team needs to learn more. The results are written up and sent to the judges who use the site visit reports and discussions with the team leaders to recommend award recipients to the Secretary of Commerce. Statistics and data analysis tools can be used to provide a summary of the examiners' scoring profiles and to help the judges review the scores. Figure below illustrates a hypothetical example (Excel file Baldrige). Your task is to apply the

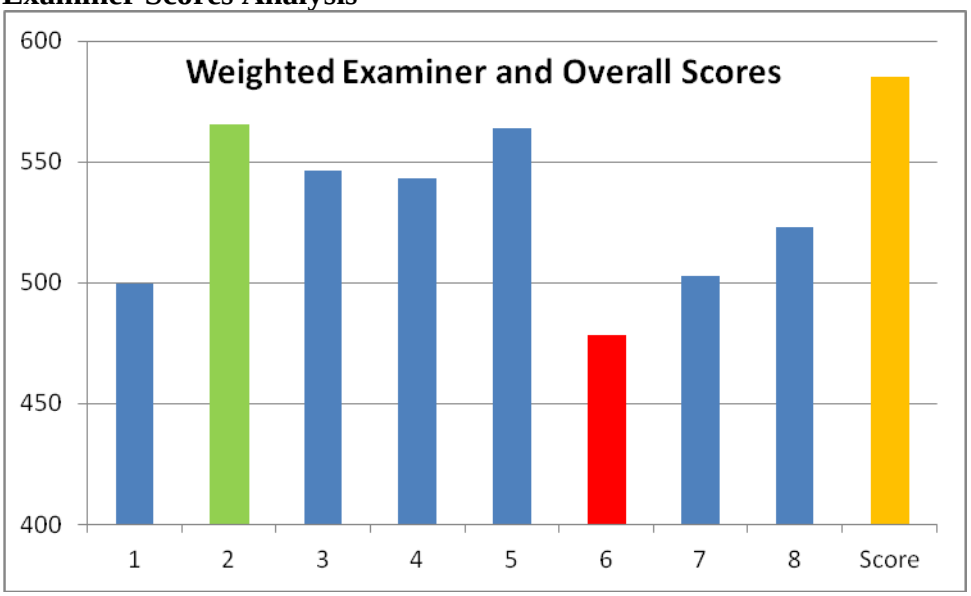
concepts and tools discussed in this chapter to analyze the data and provide the judges with appropriate statistical measures and visual information to facilitate their decision process regarding a site visit recommendation.





The highest absolute difference between the consensus and median scores is 15, few 10, but mostly the difference was within 5 points. Most of the consensus scores are higher (or the same) than the median, indicating that information sharing and discussion among examiners generally raises the consensus score, i.e. exhibits a form of positive group bias.

Examiner Scores Analysis



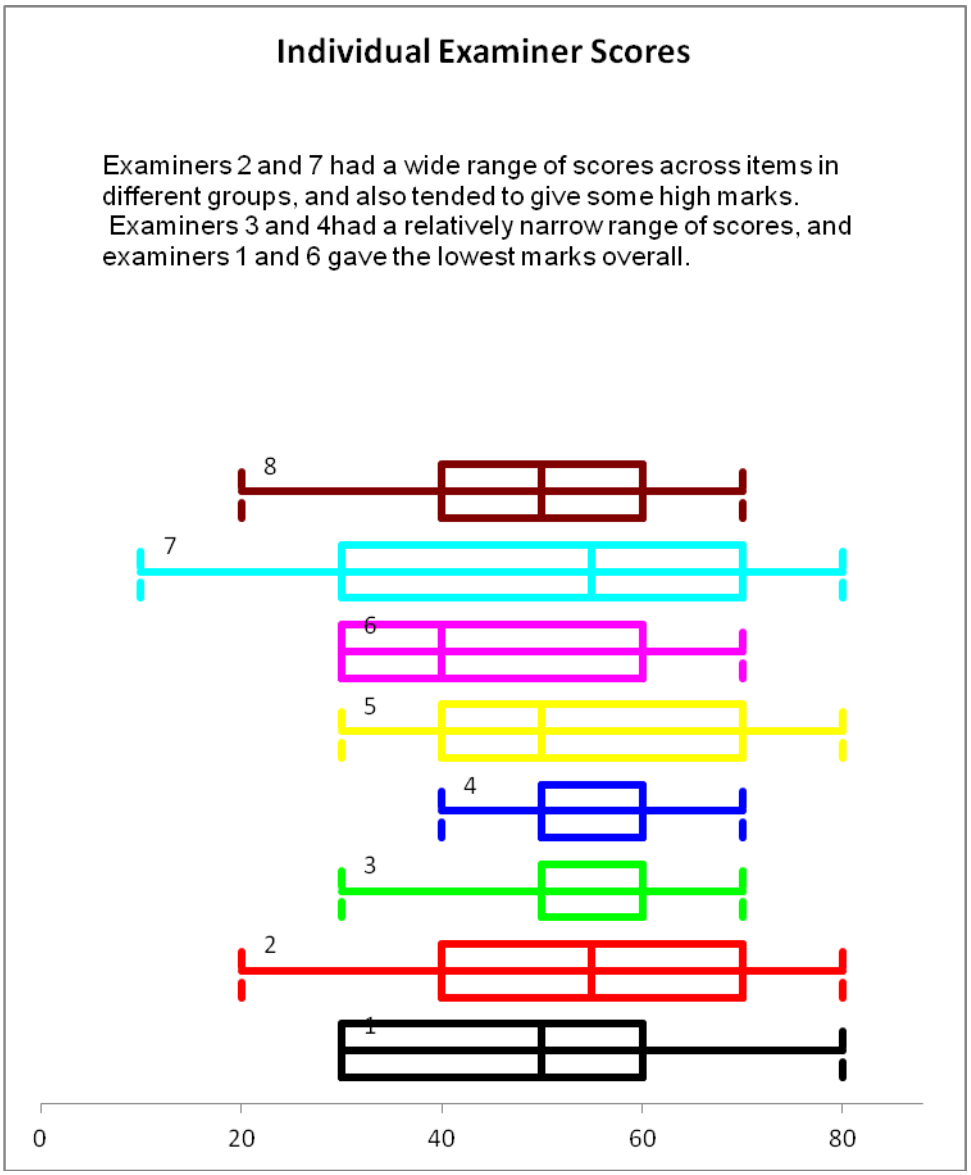
Most examiners come in the low 500's, with two in high 500's, including the overall score. Only one examiner's weighted score was below 500.

Examiner Scores Analysis

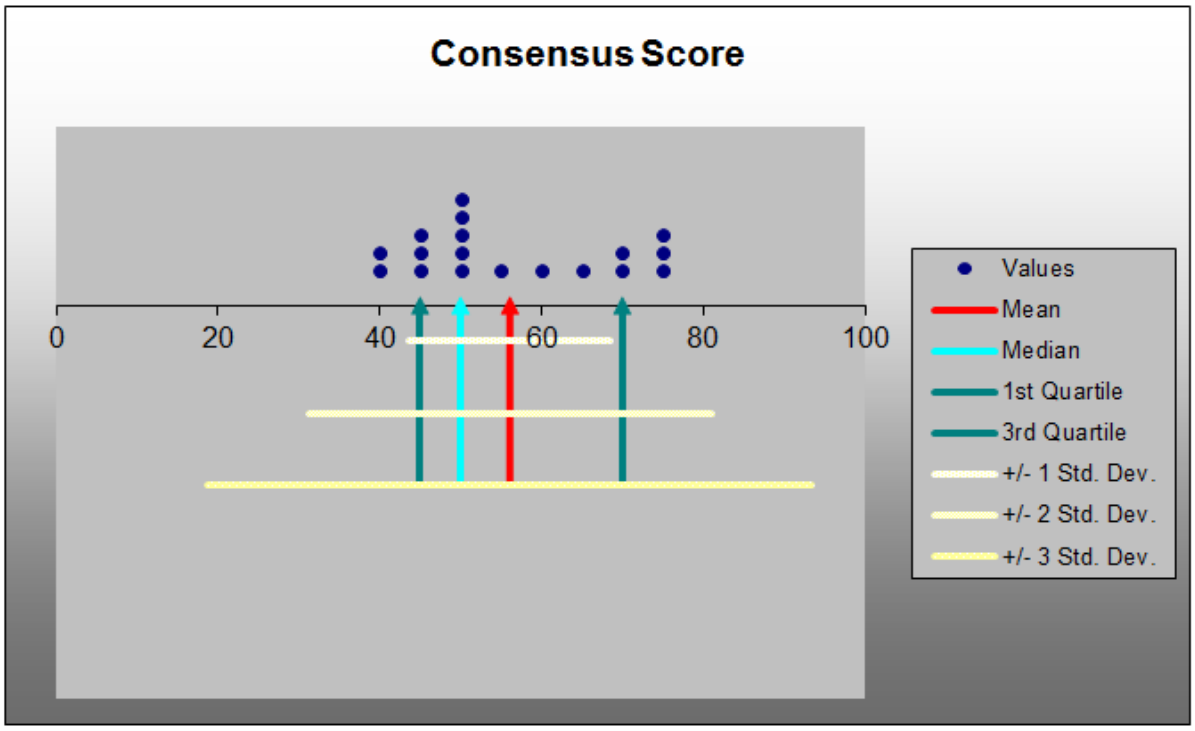
	1	2	3	4	5	6	7	8	Score
1	1.000								
2	0.524	1.000							
3	0.473	0.460	1.000						
4	0.320	0.713	0.523	1.000					
5	0.308	0.183	0.480	0.391	1.000				
6	0.598	0.489	0.255	0.257	0.196	1.000			
7	0.616	0.560	0.414	0.345	0.222	0.700	1.000		
8	0.239	0.530	0.760	0.453	0.428	0.346	0.309	1.000	
Score	0.729	0.743	0.669	0.590	0.391	0.726	0.799	0.616	1.000

Examiners were pretty consistent in their scoring of individual items, with the overall score relatively highly correlated with individual examiners, with the possible exception of examiner # 5.

Box plot for Examiner Scores



Dot chart for Examiner Scores



Summary	
Mean	56.1111
Median	50.0000
1st Quartile	45.0000
3rd Quartile	70.0000
Std. Deviation	12.4328