Assignment module one

Mayel Espino

[2] University of San Diego

Author Note

Abstract

This is the assignment exercises and answers for module one of ADS500B.

Keywords:

Assignment exercise one

The Exercise 1 Dataset (located in your assignment prompt in Blackboard) contains an imaginary dataset of auto insurance providers and their ratings as provided by the latest three customers. Now if you had to choose an auto insurance provider based on these ratings, which one would you opt for and why?

Assignment exercise two

The Exercise 2 Dataset (located in your assignment prompt in Blackboard) for this problem contains statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973. Also given is the percentage of the population living in urban areas. Use the pre-processing techniques at your disposal to prepare the dataset for analysis.

2.1: Resolve the missing values without deleting rows.

2.2: Look for outliers and smooth noisy data without deleting rows.

2.3: Prepare the dataset to establish a relation between an urban population category and a crime type. Submit your final dataset with imputed and smoothed values. [Hint: Convert the urban population percentage into categories, for example, small (<50%), medium (<60%), large (<70%), and extra-large (70% and above) urban population.]

Assignment exercise three

The Exercise 3 Dataset (located in your assignment prompt in Blackboard) for this problem is a dataset of bridges in Pittsburgh. The original dataset was prepared by Yoram Reich and Steven J. Fenves, Department of Civil Engineering and Engineering Design Research Center, Carnegie Mellon University. Use this dataset to complete the following tasks:

3.1: Look for outliers and smooth noisy data.

3.2: Resolve all the missing values without deleting rows.

3.3: Prepare the dataset to establish a relation among:

Length of the bridge and its purpose.

Number of lanes and its materials.

Span of the bridge and number of lanes.

Assignment exercise four

A community library has decided to limit all future procurement of books either to hardback or to softback copies. The library also plans to convert all the existing books to one cover type later. Fortunately, to help you decide, the library has gathered a small sample of data (Exercise 4 Dataset located in your assignment prompt in Blackboard) that gives measurements on the volume, area (only the cover of the book), and weight of 15 existing books, some of which are softback ("pb") and the rest are hardback ("hb") copies.

The Exercise 4 Dataset represents that the dataset has 15 instances of the following four attributes:

Volume: Book volumes in cubic centimeters Area: Total area of the book in square centimeters

Weight: Book weights in grams

Cover: A factor with levels; Hb for hardback, and Pb for paperback

Now use this dataset to decide which type of book you want to procure in the future. Here is how you are going to do it. Determine:

4.1: The mode of the book covers.

4.2: The mean of the book weights by book covers.

4.3: The variance in book volumes.

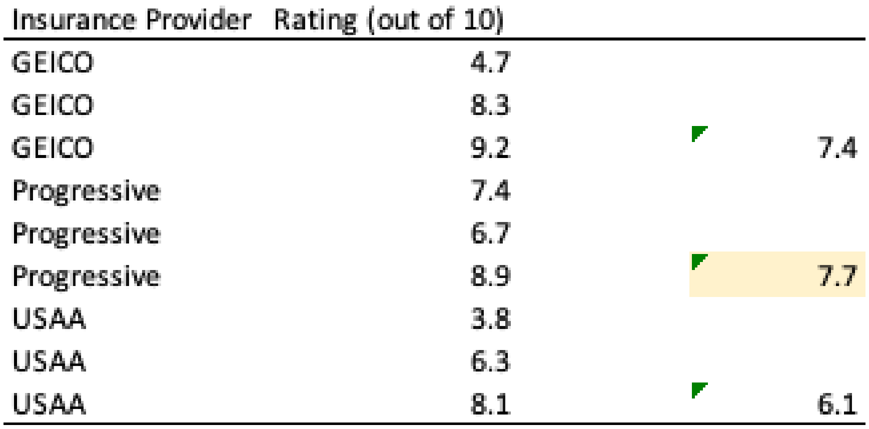
4.4: Use the above values to decide which book cover types the library should opt for in the future.

Results

Answer for assignment exercise one

Table 1

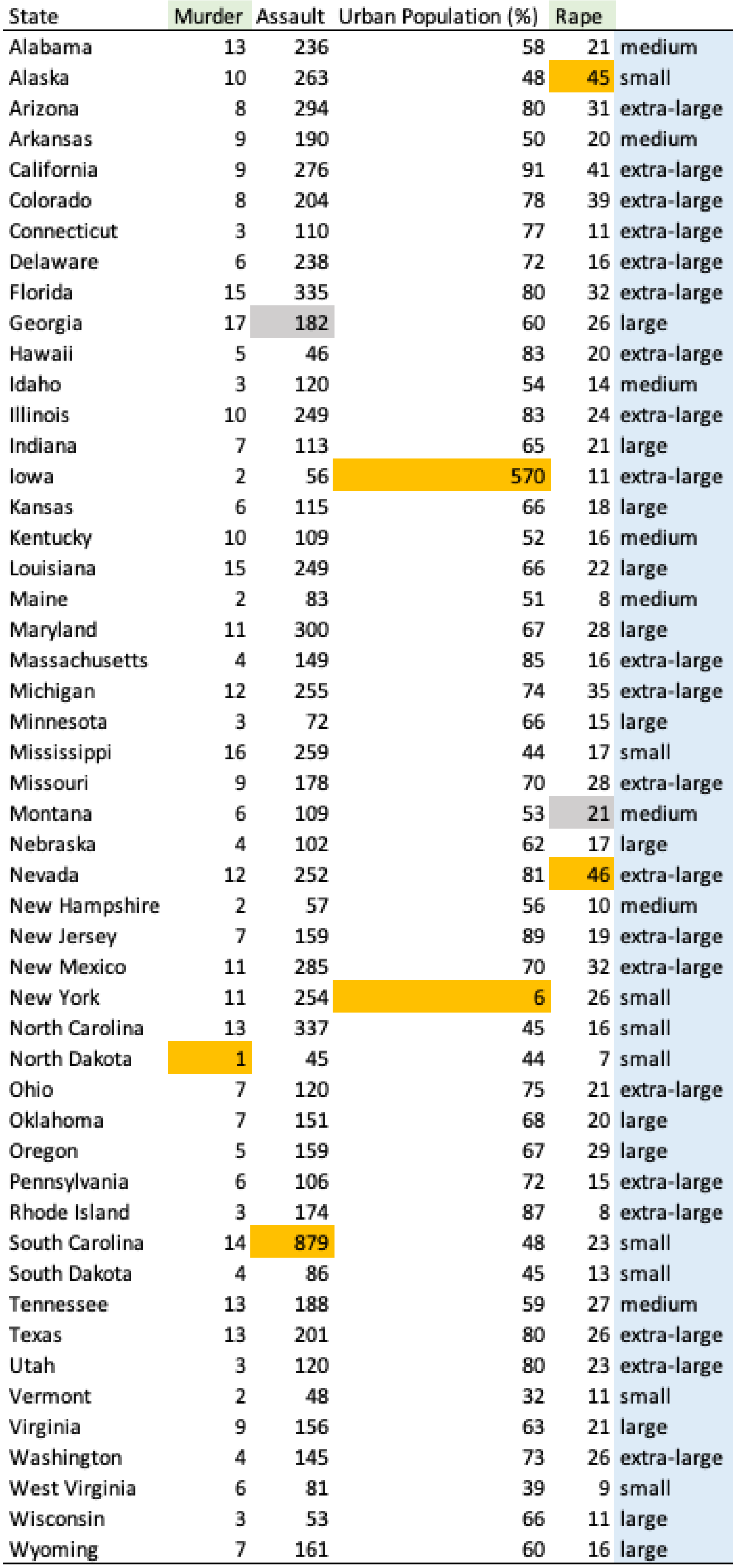
Average customer rating for Insurance providers



Insurance Provider	Rating (out of 10)	
GEICO	4.7	
GEICO	8.3	
GEICO	9.2	7.4
Progressive	7.4	
Progressive	6.7	
Progressive	8.9	7.7
USAA	3.8	
USAA	6.3	
USAA	8.1	6.1

Based on the average customer rating I would select Progressive.

Answer for assignment exercise two



State	Murder	Assault	Urban Population (%)	Rape	
Alabama	13	236	58	21	medium
Alaska	10	263	48	45	small
Arizona	8	294	80	31	extra-large
Arkansas	9	190	50	20	medium
California	9	276	91	41	extra-large
Colorado	8	204	78	39	extra-large
Connecticut	3	110	77	11	extra-large
Delaware	6	238	72	16	extra-large
Florida	15	335	80	32	extra-large
Georgia	17	182	60	26	large
Hawaii	5	46	83	20	extra-large
Idaho	3	120	54	14	medium
Illinois	10	249	83	24	extra-large
Indiana	7	113	65	21	large
Iowa	2	56	570	11	extra-large
Kansas	6	115	66	18	large
Kentucky	10	109	52	16	medium
Louisiana	15	249	66	22	large
Maine	2	83	51	8	medium
Maryland	11	300	67	28	large
Massachusetts	4	149	85	16	extra-large
Michigan	12	255	74	35	extra-large
Minnesota	3	72	66	15	large
Mississippi	16	259	44	17	small
Missouri	9	178	70	28	extra-large
Montana	6	109	53	21	medium
Nebraska	4	102	62	17	large
Nevada	12	252	81	46	extra-large
New Hampshire	2	57	56	10	medium
New Jersey	7	159	89	19	extra-large
New Mexico	11	285	70	32	extra-large
New York	11	254	6	26	small
North Carolina	13	337	45	16	small
North Dakota	1	45	44	7	small
Ohio	7	120	75	21	extra-large
Oklahoma	7	151	68	20	large
Oregon	5	159	67	29	large
Pennsylvania	6	106	72	15	extra-large
Rhode Island	3	174	87	8	extra-large
South Carolina	14	879	48	23	small
South Dakota	4	86	45	13	small
Tennessee	13	188	59	27	medium
Texas	13	201	80	26	extra-large
Utah	3	120	80	23	extra-large
Vermont	2	48	32	11	small
Virginia	9	156	63	21	large
Washington	4	145	73	26	extra-large
West Virginia	6	81	39	9	small
Wisconsin	3	53	66	11	large
Wyoming	7	161	60	16	large

Table 2

statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973

Table 2 is the result of processing done to the data set provided. The following are the steps taken:

For columns murder and rape, I removed all the decimal points.

For columns assault and rape, to find the missing values without removing them, I calculated the average for each column. The value I replaced is highlighted in grey.

To identify the outliers, highlighted in orange, I compared the extreme values to the average and the frequency of similar values. So for example Murder has an outlier value of 1, while most of the smaller values are 2 and 3.

The following are the outliers per column:

Murder outlier: 1

Assault outlier: 879

Urban population outliers: 570 and 6

Rape outliers: 45 and 46

To support my findings, below are the boxplots for each of columns:

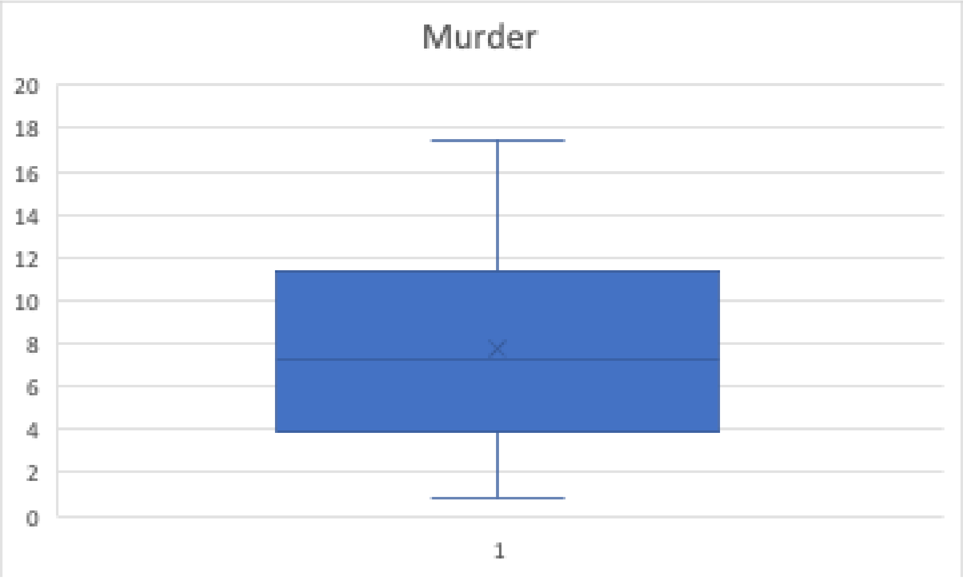


Figure 1

Murder boxplot

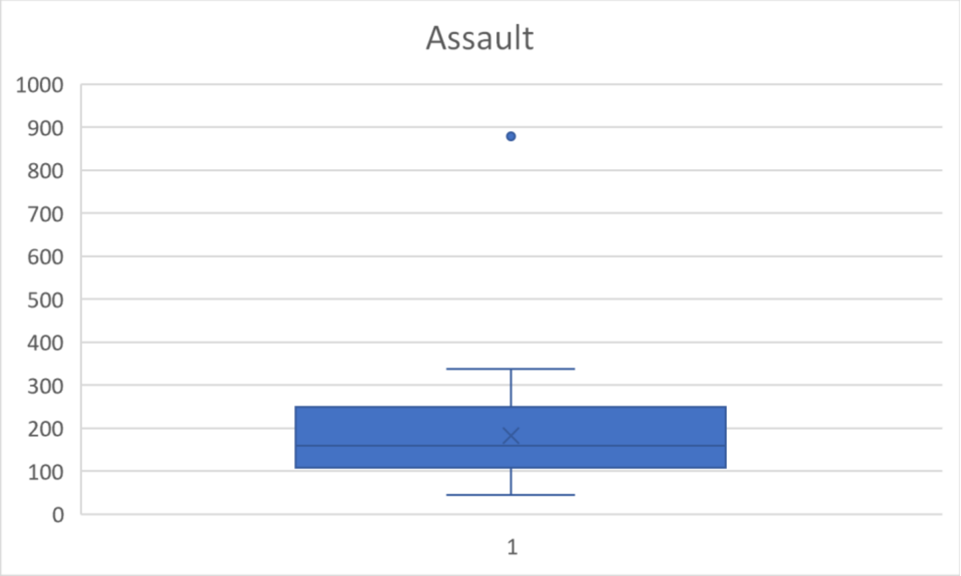


Figure 2

Assault boxplot.

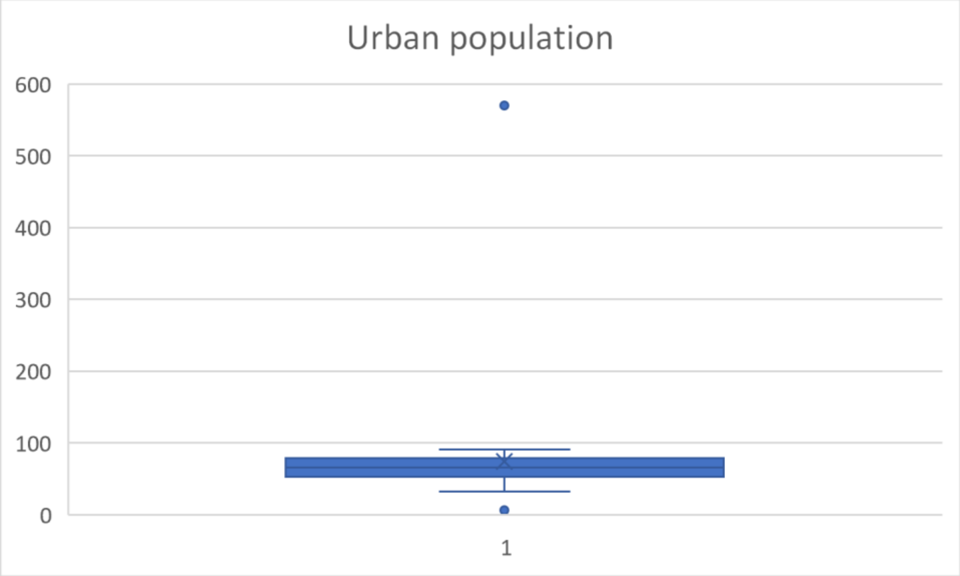


Figure 3

Urban population boxplot

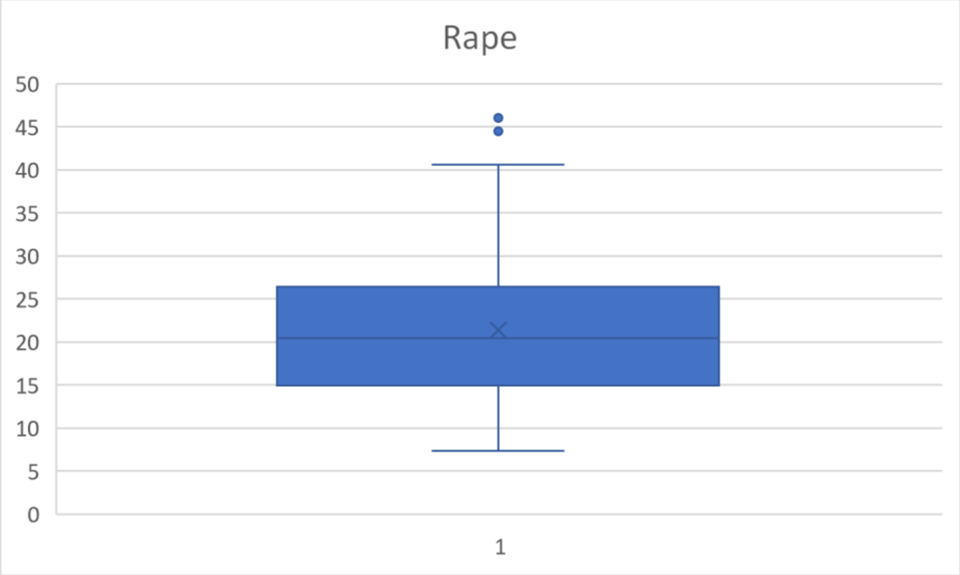


Figure 4

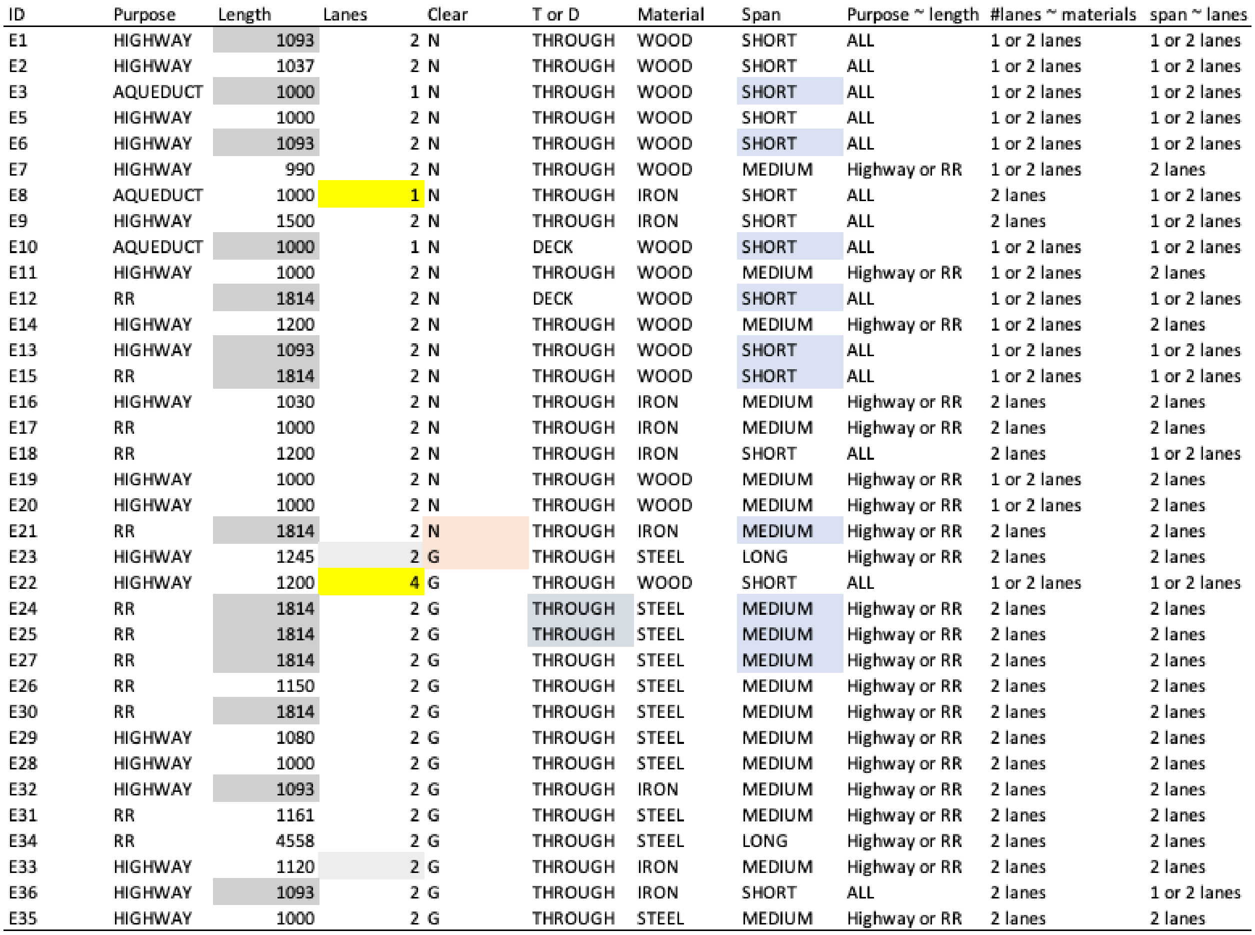
Rape boxplot

Finally the column highlighter in blue is the categorization to establish a relation between an urban population category and a crime type. This was done as per the instructions in the assignment, categorizing the populations in to small, large and extra-large.

Answer for assignment exercise three

Table 3

Bridges in Pittsburgh



ID	Purpose	Length	Lanes	Clear	T or D	Material	Span	Purpose ~ length	#lanes ~ materials	span ~ lanes
E1	HIGHWAY	1093	2	N	THROUGH	WOOD	SHORT	ALL	1 or 2 lanes	1 or 2 lanes
E2	HIGHWAY	1037	2	N	THROUGH	WOOD	SHORT	ALL	1 or 2 lanes	1 or 2 lanes
E3	AQUEDUCT	1000	1	N	THROUGH	WOOD	SHORT	ALL	1 or 2 lanes	1 or 2 lanes
E5	HIGHWAY	1000	2	N	THROUGH	WOOD	SHORT	ALL	1 or 2 lanes	1 or 2 lanes
E6	HIGHWAY	1093	2	N	THROUGH	WOOD	SHORT	ALL	1 or 2 lanes	1 or 2 lanes
E7	HIGHWAY	990	2	N	THROUGH	WOOD	MEDIUM	Highway or RR	1 or 2 lanes	2 lanes
E8	AQUEDUCT	1000	1	N	THROUGH	IRON	SHORT	ALL	2 lanes	1 or 2 lanes
E9	HIGHWAY	1500	2	N	THROUGH	IRON	SHORT	ALL	2 lanes	1 or 2 lanes
E10	AQUEDUCT	1000	1	N	DECK	WOOD	SHORT	ALL	1 or 2 lanes	1 or 2 lanes
E11	HIGHWAY	1000	2	N	THROUGH	WOOD	MEDIUM	Highway or RR	1 or 2 lanes	2 lanes
E12	RR	1814	2	N	DECK	WOOD	SHORT	ALL	1 or 2 lanes	1 or 2 lanes
E14	HIGHWAY	1200	2	N	THROUGH	WOOD	MEDIUM	Highway or RR	1 or 2 lanes	2 lanes
E13	HIGHWAY	1093	2	N	THROUGH	WOOD	SHORT	ALL	1 or 2 lanes	1 or 2 lanes
E15	RR	1814	2	N	THROUGH	WOOD	SHORT	ALL	1 or 2 lanes	1 or 2 lanes
E16	HIGHWAY	1030	2	N	THROUGH	IRON	MEDIUM	Highway or RR	2 lanes	2 lanes
E17	RR	1000	2	N	THROUGH	IRON	MEDIUM	Highway or RR	2 lanes	2 lanes
E18	RR	1200	2	N	THROUGH	IRON	SHORT	ALL	2 lanes	1 or 2 lanes
E19	HIGHWAY	1000	2	N	THROUGH	WOOD	MEDIUM	Highway or RR	1 or 2 lanes	2 lanes
E20	HIGHWAY	1000	2	N	THROUGH	WOOD	MEDIUM	Highway or RR	1 or 2 lanes	2 lanes
E21	RR	1814	2	N	THROUGH	IRON	MEDIUM	Highway or RR	2 lanes	2 lanes
E23	HIGHWAY	1245	2	G	THROUGH	STEEL	LONG	Highway or RR	2 lanes	2 lanes
E22	HIGHWAY	1200	4	G	THROUGH	WOOD	SHORT	ALL	1 or 2 lanes	1 or 2 lanes
E24	RR	1814	2	G	THROUGH	STEEL	MEDIUM	Highway or RR	2 lanes	2 lanes
E25	RR	1814	2	G	THROUGH	STEEL	MEDIUM	Highway or RR	2 lanes	2 lanes
E27	RR	1814	2	G	THROUGH	STEEL	MEDIUM	Highway or RR	2 lanes	2 lanes
E26	RR	1150	2	G	THROUGH	STEEL	MEDIUM	Highway or RR	2 lanes	2 lanes
E30	RR	1814	2	G	THROUGH	STEEL	MEDIUM	Highway or RR	2 lanes	2 lanes
E29	HIGHWAY	1080	2	G	THROUGH	STEEL	MEDIUM	Highway or RR	2 lanes	2 lanes
E28	HIGHWAY	1000	2	G	THROUGH	STEEL	MEDIUM	Highway or RR	2 lanes	2 lanes
E32	HIGHWAY	1093	2	G	THROUGH	IRON	MEDIUM	Highway or RR	2 lanes	2 lanes
E31	RR	1161	2	G	THROUGH	STEEL	MEDIUM	Highway or RR	2 lanes	2 lanes
E34	RR	4558	2	G	THROUGH	STEEL	LONG	Highway or RR	2 lanes	2 lanes
E33	HIGHWAY	1120	2	G	THROUGH	IRON	MEDIUM	Highway or RR	2 lanes	2 lanes
E36	HIGHWAY	1093	2	G	THROUGH	IRON	SHORT	ALL	2 lanes	1 or 2 lanes
E35	HIGHWAY	1000	2	G	THROUGH	STEEL	MEDIUM	Highway or RR	2 lanes	2 lanes

Steps taken to complete the assignment:

To replace the empty values in Length, I used the average of all the lengths, highlighted in grey.

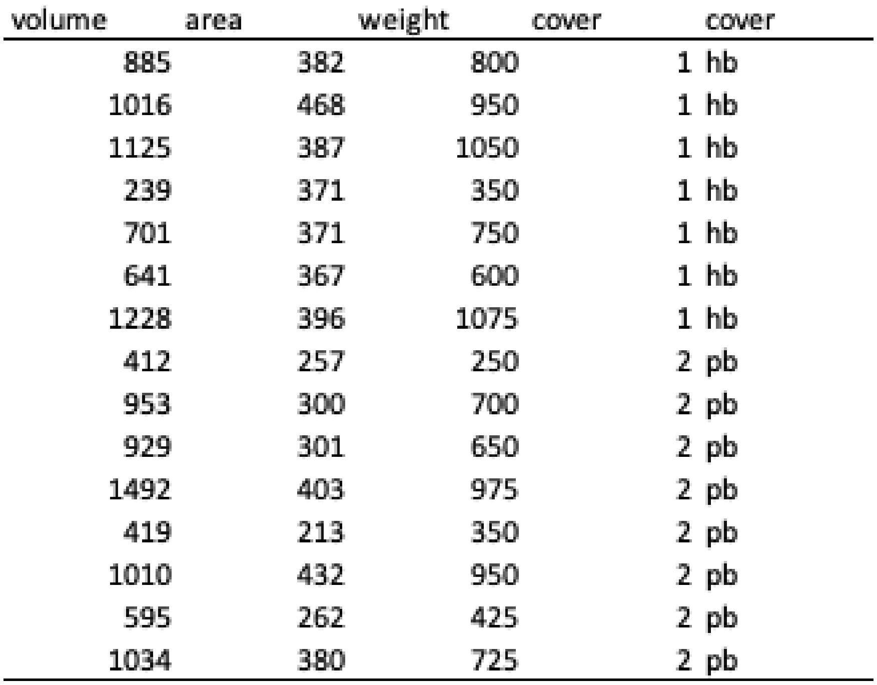
To replace the empty values in Lanes, I used the most common value of 2, highlighted in light grey.

To replace the empty values in Clear, since G and N are evenly distributed, I replaced the blank values following the same distribution, since there were two blanks, I used one G and one N. Highlighted in pink.

To replace the empty values in column "T or D" I used the most common value of TRHOUGH, highlighted in blue-grey.

To replace the empty values in span, I used the most common value based on each category: I replaced all blanks for wooden spans with short and steel and iron with medium which occurs more often than does long.

Answer for assignment exercise four



volume	area	weight	cover	cover
885	382	800	1	hb
1016	468	950	1	hb
1125	387	1050	1	hb
239	371	350	1	hb
701	371	750	1	hb
641	367	600	1	hb
1228	396	1075	1	hb
412	257	250	2	pb
953	300	700	2	pb
929	301	650	2	pb
1492	403	975	2	pb
419	213	350	2	pb
1010	432	950	2	pb
595	262	425	2	pb
1034	380	725	2	pb

Table 4

Sample data for library

Answer 1: I assigned 1 to SB and 2 to PB. The mode of book covers is PB, or 2. Based on the exploratory statistics function in Excel.

Answer 2: The mean for HB is 796.429 and for PB is 628.125. Based on the exploratory statistics function in Excel.

Answer 3: The sample variance is 115518.3524, I used the VAR function in Excel.

Answer 4: The answer to this assignment is my recommendation: buy more hard cover books. My answer is based on the fact that the mean for total area of hard cover books is 796.429 and the mean for total area of 628.125. So if there has to be a choice between the two we choose the type of cover which is most common in the population. The fact that hard cover is more common can obey to a couple of factors: popularity or durability. In either case it is a good option to select hard cover.