Chapter 1: The Roles of Data and Predictive Analytics in Business

Answers to Questions and Problems

1. a. Structured. The unit of observation is an individual, and for each one, we can identify their: name, age, height, and location clearly.

   b. Unstructured. There are clearly separate pieces of information being collected (texts, dates, prices) but there isn't a clear way to assemble them into distinct units of observation.

   c. Structured. The unit of observation is some sort of rectangular object (it need not have an explicit label for it to be well defined), and the even columns provide information on color, length, weight, and width for each one.

   d. Unstructured. There are clearly separate pieces of information (time, price, sales) but no clear way to assemble them into distinct units of observation.

2. a. The unit of observation is a store-year. Note that it may be tempting to claim the unit of observation is a store-person-year, since there is also variation in managers. However, as the data are presented, knowing the year and store automatically implies the manager; therefore, the unit of observation is a store-year only. These are panel data.

   b. The unit of observation is a month. These are time series data.

   c. The unit of observation is a person-year. These data are a pooled cross-section.

   d. The unit of observation is a factory. These are cross-sectional data.

3. a. Query. This may seem like pattern discovery, but there needs to be some threshold that qualifies this as a pattern.

   b. Query. This is simply a request for information for the dataset.

   c. Causal inference. This describes the causal effect of advertising on sales.

   d. Pattern discovery. This is a form of outlier detection.

   e. Pattern discovery. This is a form of association analysis.

4. Lead information pertains to what will happen and lag information pertains to what did happen.

5. Passive prediction involves predicting outcomes while observing, but not altering, their determining factors. Active prediction involves predicting outcomes after altering at least one of their determining factors.

6. a. Active prediction – Tom exogenously alters his diet.

   b. Passive prediction – Ann does not directly alter the number of visits to her site.

   c. Active prediction – Laura exogenously alters her advertising.

   d. Passive prediction – Alex does not directly alter people's credit card purchasing.

   e. Passive prediction – John does not directly alter the voter's answers.

7. As stated in the text, it allows the decision-maker to make evidence-based assessments of expected outcomes from alternative strategies, and then choose the optimal one based on her business objective.

8. a. i. A theoretical refute may be as simple as follows. Based on your own sense of the matter, people find the ads entertaining, but not enough to substantially respond in terms of purchasing. Therefore, the change in sales will not be enough to offset the costs of the ads, meaning increased ad expenditure will lower profits.

      ii. You collect data on varying levels of ad expenditure along with profits across locations and/or time. Then, using techniques described in later chapters, you analyze how profits respond to changes in ad expenditure in the data. If the analysis shows profits declining with increases in ad expenditure, this would constitute a refutation to the claim.

   b. Data consist of what actually occurred, allowing for evidence-based decision-making, rather than "gut"-based decision-making.

9. Here, we need three factors that we believe have a causal effect on the number of years an employee stays with a firm. Three such factors might be:

   1. Education (which might influence the employee's competing options)

2. Number of nearby rival firms (also may influence the employee's competing options)
3. Age when hired (which may be indicative of the employee's job mobility)

10. Following the example in the text, we can formally express the data generating process for weekly soda sales as: $Sales_t = f(Price_t, Placement_t, Holiday_t) + U_t$.

11. a. This does not require active prediction. Rather, it is a good example of an application of passive prediction. We want to predict how purchases relate to age, and we are not making changes to our customers' ages.

    b. This does require active prediction. We are considering making an active change in strategy – in the form of a new celebrity endorsement – and we want to predict how sales will respond.

    c. This does require active prediction. We are actively changing product placement (a strategic move), and want to know the impact on profits.

12. Amanda is making the active prediction. She is determining what will happen with a change in strategy (i.e., a price cut). In comparison, Darryl is making a passive prediction. He is using demographics – which Meredith is not considering, or capable of, changing – to predict the likelihood of an accident.

13. See *DataLoad.xlsx* for the data loading, or the table below provides an example.

Name	Year	Vote	Score
Laurene Horton	2015	Yes	47
Wilson Zimmerman	2015	No	83
Jeffrey Wade	2015	No	26
Candice Graves	2015	Yes	91
Kayla Snyder	2015	Yes	52
Laurene Horton	2016	No	83
Wilson Zimmerman	2016	No	76
Jeffrey Wade	2016	No	35
Candice Graves	2016	Yes	62
Kayla Snyder	2016	No	48

The unit of observation is a person-year. The data are panel data.

14. (a,b,c): See *Scorecard Answer.xlsx*

Goals	Region	Measure	Target	Result	Performance
Revenues	North	Average Revenue	$200,000	$183,913.45	Too Low
	South			$204,704.86	Good
	East			$204,976.64	Good
Growth	North	Average Growth	5%	3.78%	Too Low
	South			3.83%	Too Low
	East			4.67%	Almost
Returns	North	Maximum Store Return	$10,000	$10,202	Too High
	South			$10,280	Too High
	East			$10,297	Too High

15. a. i. $1,468,424.42
    ii. 11,526,750.78
    iii. 3,579,884.506
    iv. $51,439.46
    v. $353,890,286.00

    b. Two candidates include mean of Materials Costs ($21,161.41) and Variance of Labor Costs ($87,892,572).

16. a. $1,481,100.02
    b. $504,655
    c. Region 166
    d. Region 223
    e. $2,397,435 - $500,776 = $1,896,659

17. a. There is a strong positive correlation between a customer being active and their age level. Hence, it appears younger customers are most likely to drop. This is lead information since it is designed to look ahead and assess where the greatest risks of customer loss will be in the future.

    b. The North Region had the most customers (84). This is lag information, since it is simply reporting what happened.