Solutions Manual – Chapter 2

**Solutions to Discussion Questions**

1. The information needs only to be entered once and changes or edits only need to be done in one file versus multiple files. It won’t take up unnecessary space (which is expensive), take up unnecessary processing to run reports to ensure that there aren’t multiple versions of the truth, and will not increase the risk of data entry errors.
2. Relational databases are designed to support business processes across the organization, which results in improved communication across functional areas and more integrated business processes.
3. Relational databases all connect with each other by use of the primary and foreign key. That makes data analysis very easy to do since you can readily join the tables and run the requested data analysis.
4. Relational databases can be designed to aid in the placement and enforcement of internal controls and business rules in ways that flat files cannot. Due to the nature of the primary key/foreign key, both a primary key and a foreign key must line up with each other before any business can be transacted. If there is no supplier in the approved supplier file, it is not possible to process a purchase order without linking to the approved supplier file.
5. The data dictionary is a centralized repository of descriptions for all of the data attributes of the data set. Attributes of a data dictionary for each field might include a variable name, a brief description, whether the field is made up of numbers or text or alphanumerics, the size (or number of digits) of the field, whether it serves as a primary or foreign key and notes, etc.
6. Before extracting the data, it is important to be able to answer these questions:
   1. What is the purpose of the data request? What do you need the data to solve? What business problem will it address?
   2. What risk exists in data integrity (e.g., reliability, usefulness)? What is the mitigation plan?
   3. What other information will impact the nature, timing and extent of the data analysis?
7. The analyst needs to know what data is available, how it comes, what it includes, and how reliable the data is to be able to answer the central question which was the reason for the analysis.
8. The more frequent the requested report, the more the database administrator will set it up for automatic extraction and delivery. It may also be a question of how often the data changes. If the data is updated weekly and the data is extracted daily, that may not make any sense.
9. The database administrator is most familiar with the data and may be able to help the analyst get the data needed to address the question. There also might be some sensitivities to who gets what data to ensure that the data gets to the intended analyst and audience.
10. The impact of transforming data to work with NULL, N/A and zero values in the dataset might have an impact on programs like Excel.
    1. Transforming NULL and N/A values into blanks.
       1. The COUNT and AVERAGE functions would not include these fields in their computation for these variables.
    2. Transforming NULL and N/A values into zeroes.
       1. The COUNT and AVERAGE functions would incorporate these zeroes and would be included in their computation for these variables. It would have an impact particularly on the computation of the average since it would have the value of zero.
    3. Deleting records that have NULL and N/A values from your dataset.
       1. The COUNT and AVERAGE functions would not include these fields in their computation for these variables. If they are deleted all of the other fields and variables would be deleted as well, thus having a bigger impact on the overall dataset.

**Solutions to Problems**

Problem 2.1

Attributes needed from the College Scorecard data to compare the cost of attendance across types of institutions (public, private non-profit, private for-profit) would include:

* 1. CONTROL – 1 = Public. 2 = Private nonprofit. 3 = Private for-profit
  2. COSTT4\_A – Average cost of attendance

Problem 2.2

Attributes needed from the College Scorecard data to compare SAT scores across types of institutions (public, private non-profit, private for-profit) would include:

1. CONTROL – 1 = Public. 2 = Private nonprofit. 3 = Private for-profit
2. SAT\_AVG – average equivalent SAT of students admitted
3. UNITID – a unique identifier for the institution

Problem 2.3

Attributes needed from the College Scorecard data to compare levels of diversity across types of institutions (public, private non-profit, private for-profit) would include:

1. CONTROL – 1 = Public. 2 = Private nonprofit. 3 = Private for-profit
2. UNITID – a unique identifier for the institution
3. UGDS – enrollment of undergraduate certificate/degree-seeking students
4. UGDS\_WHITE – total share of enrollment of undergraduates who are white
5. UGDS\_BLACK – total share of enrollment of undergraduates who are black
6. UGDS\_HISP – total share of enrollment of undergraduates who are Hispanic
7. UGDS\_ASIAN – total share of enrollment of undergraduates who are Asian
8. UGDS\_AIAN – total share of enrollment of undergraduates who are American Indian/Alaska Native
9. UGDS\_NHPI – total share of enrollment of undergraduates who are Native Hawaiian/Pacific Islander
10. UGDS\_2MOR – total share of enrollment of undergraduates who are two or more races
11. UGDS\_NRA – total share of enrollment of undergraduates who are non-resident aliens
12. UGDS\_UNKN – total share of enrollment of undergraduates whose race is unknown

Problem 2.4

Attributes needed from the College Scorecard data to compare completion rate across types of institutions (public, private non-profit, private for-profit) would include:

1. CONTROL – 1 = Public. 2 = Private nonprofit. 3 = Private for-profit
2. UGDS – enrollment of undergraduate certificate/degree-seeking students
3. UNITID – a unique identifier for the institution

Problem 2.5

Attributes needed from the College Scorecard data to compare the percentage of students who receive federal loans at universities above and below the median cost of attendance across all institutions (public, private non-profit, private for-profit) would include:

1. CONTROL – 1 = Public. 2 = Private nonprofit. 3 = Private for-profit
2. PCTFLOAN – Percent of all federal undergraduates receiving a federal student loan
3. COSTT4\_A – Average cost of attendance
4. UNITID – a unique identifier for the institution

Problem 2.6

Attributes needed from the College Scorecard data to compare the percentage of students who receive federal loans at universities above and below the median cost of attendance across all institutions (public, private non-profit, private for-profit) would include:

1. UNITID – a unique identifier for the institution
2. STABBR – State postcode
3. COSTT4\_A – Average cost of attendance

Problem 2.7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Section 1: To be completed by customer/requestor | | | | |
| Name | Vernon Richardson | Contact Number | 479-345-0537 | |
| Email Address | vjricha@uark.edu | |
| **Description of Information Required** *(Please include dates/timeframes for any analysis, and other specific indicators/categories required in the data)* | From the College Scorecard data, I need the following data items for each year and unique identifier (UNITID) of the dataset:   1. UNITID – a unique identifier for the institution 2. STABBR – State postcode 3. COSTT4\_A – Average cost of attendance | | | |
| **Purpose/Context** *(what the data is required for)* | I am trying to determine if different regions of the country have significantly different costs of attendance. | | | |
| **Frequency** *(circle as appropriate)* | *One Off Request / Other* | *Annually*  *X* | | *Term* |
| Request Date | 7/15/2020 | Required Date | | 7/28/2020 |
| **Format Required** *(Table, Spreadsheet, Word, etc.) – please specify* | Tab Delimited File | **Customer** *(if not requestor)* | |  |
| **To be used in** *(presentation, report, etc.) – please specify* | Report to University of Arkansas administration | **Intended Audience** *(if appropriate)* | | Chancellor, President, Provost of the University of Arkansas |

Problem 2.8

Diversity can be determined by a number of different dimensions. The College Scorecard data seems to have information on race including the following fields:

UGDS – enrollment of undergraduate certificate/degree-seeking students

UGDS\_WHITE – total share of enrollment of undergraduates who are white

UGDS\_BLACK – total share of enrollment of undergraduates who are black

UGDS\_HISP – total share of enrollment of undergraduates who are Hispanic

UGDS\_ASIAN – total share of enrollment of undergraduates who are Asian

UGDS\_AIAN – total share of enrollment of undergraduates who are American Indian/Alaska Native

UGDS\_NHPI – total share of enrollment of undergraduates who are Native Hawaiian/Pacific Islander

UGDS\_2MOR – total share of enrollment of undergraduates who are two or more races

UGDS\_NRA – total share of enrollment of undergraduates who are non-resident aliens

UGDS\_UNKN – total share of enrollment of undergraduates whose race is unknown

Depending on the focus of the report, it may make sense to capture broader dimensions of diversity rather than knowing the population of each individual race category. In that case, it may make sense to combine categories to have less categories.

Problem 2.9

You would first need to calculate the median cost of attendance at universities and determine which universities are above and below that median. You may need to do this for each year included in the analysis as the cost of attendance changes from year to year. Once this is done, you can compute the percentage of students who receive federal loans at each university and compare them for those both above and below the median cost of attendance.