Course Learning Outcomes for Unit VIII

Upon completion of this unit, you should be able to:

1. Analyze techniques used for data collection.

2. Summarize the steps for connecting data to a business intelligence (BI) environment.

4. Determine the access controls for data visualizations.

6. Evaluate data analytic outputs for decision-making.
   6.1 Determine the best approach for analyzing outputs for a department's performance.
   6.2 Examine the process of reviewing missing data.

Required Unit Resources

Chapter 9: Telling the Truth with Data Visualization
It is not required to read the Glossary and Problems sections at the end of the chapter.

Unit Lesson

In our previous units, we described creating compelling visualizations as storytelling. There is a hidden gem of honesty in the noun *storytelling* that may not be immediately obvious. When we tell stories, we generally want to influence our audience—often to create a sense of wonder, mystery, joy, tragedy, or compassion. An author will selectively choose which details to include in a story to develop the desired reaction.

Think of the embellished tales you have heard about how great an athlete someone was in high school or the giant sailfish they caught during their last fishing expedition. Details are left out, taken out of context, or the audience's focus is drawn to a singular aspect, while other information is purposefully diminished. Just because the storyteller ultimately creates the desired influence on their audience does not mean they are being accurate or honest. That happens with data visualization also.

Storytelling With Data

Charts can be constructed to draw the audience's attention to an aspect of the data that supports the developer's goals while ignoring other parts of the same information that might lead an audience member to a contrary impression or conclusion. We must avoid telling erroneous, misleading, or incomplete stories with our visualizations. How to do so—and thereby become an impartial, accurate, and honest storyteller—is appropriately the subject of this final unit. As business intelligence developers, we must treat the data ethically and ensure that visualizations are accurate and reflect the true picture. There are times when we do need to remove transactions or entries, but these actions need to be documented and approved by the business.

Data visualizations can be inaccurate when the data they are predicated upon is incomplete. Datasets must therefore be analyzed to determine if there are any missing values before attempting to create a data visualization. Missing values can badly skew the results of a computation. It is appropriate to assess whether some value can be substituted into the cell where data is missing based on available information. Again, this needs to be a conversation with the business as opposed to having business intelligence developers making at-will decisions based on their experiences.

Data can also be erroneously entered or calibrated or otherwise contain some type of error. There are various tools at one's disposal to discover outliers and other quality issues, such as frequency distributions, scatter charts, or summary stats. Again, these types of errors can result in misleading results.

Partial Data and Selection Bias

Partial data is sample data that is not representative of the overall population. This can be due to what is known as *selection bias*, meaning that the sample was not properly randomized. For instance, if someone is conducting a study on whether Americans favor dismissal of student loans, but their selection only comes from college campus polls, there is likely to be a biased result where loan dismissal initiatives are more preferred by those polled than they would be if others who had not gone to college and amassed student loan debt had also been polled.

Survivor Bias

A similar concept is known as survivor bias. *Survivor bias* occurs when a sample dataset contains too many observations correlating to positive outcomes for an event. This condition would again result from the sample not being representative of the entire population. For instance, suppose data is being collected on the most significant causes of stress on college campuses, but only those who had fully completed their degree plan and graduated were polled. The results may indicate that the most important cause of stress on college campuses is debt or getting a first job. However, if others who were still undergoing their course of study or who had dropped out had been asked, they might have responded quite differently. Because we have no data from these students, we cannot make a proper conclusion based on the sample data.

Inflation and Price Index

Let's face it—money makes the world go round. For this reason, many visualizations are concerned with money, earnings, costs, profits, and potential revenue. However, prices change over time, so comparing cash from one point in time to another can be misleading. *Inflation* is the term that describes the tendency of things to cost more and more as time passes. Inflation can be measured by analyzing and comparing the costs of products and services, known as the *price index*. There is not one set index for all consumption and production; there are many. Therefore, it is necessary to determine which index is most closely related to the product being analyzed to appropriately take inflation into account and not merely rely on nominal values that do not reflect inflation.

Geographical Impact on Data

Another area in which one must not create misleading data visualizations relates to differences across geographical regions. This is especially true where disparities can be easily explained by population density rather than an actual difference in the analyzed factor. For instance, if one were making a visualization with an aim to determine which state has the most productive workforce that creates widgets, it would be misleading to chart only the total number of widgets produced per year in each state irrespective of the total number of workers. In such a comparison, the total widget output of California (the most populous state in the United States) would easily eclipse the total production output of Wyoming (the least populated state in the United States), even though, in reality, each worker may be producing the same number of widgets per year, and the productivity of individual workers in each state is therefore actually equal.

In Conclusion

When creating data visualization, our goal is always to tell a story to our audience that helps them understand, analyze, and problem-solve according to their needs. But we must be sure that we are telling a complete, accurate, and objectively true story when we engage in this type of storytelling. To do so, we must take into account whether our data is incomplete or erroneous, whether our data samples are affected by some form of bias, whether our financial data properly takes into account the inflationary reality of costs between different periods, and whether geographic disparities have been adequately accounted for. We can create more complete, helpful, and accurate data visualizations by addressing these common issues.

References

Killiam, G. [GarryKillian]. (n.d.). *Vector abstract colorful financial big data graph visualization* [Image]. Freepik. https://www.freepik.com/free-vector/market-research-isometric-illustration_17714187.htm#query=data%20analytics%20business%20intelligence&position=15&from_view=search&track=ais

macrovector. (n.d.). *Market research isometric illustration* [Image]. Freepik. https://www.freepik.com/free-vector/market-research-isometric-illustration_17714187.htm#query=data%20analytics%20business%20intelligence&position=15&from_view=search&track=ais

Learning Activities (Nongraded)

Nongraded Learning Activities are provided to aid students in their course of study. You do not have to submit them. If you have questions, contact your instructor for further guidance and information.

It is highly recommended that you complete the problems at the end of Chapter 9. Doing so will help you gain a better understanding of the topics taught in the unit.