Executive Summary of Research Assessment #9

As I begin to work on my Final Product for ISM, I wanted to expand my knowledge on data visualization as it will be the main focus of my project. This research assessment covers my research on a subtopic of data visualization: data cleansing - an essential aspect in creating a visualization.

Research Assessment #9

Date: February 11, 2021

Subject: Understanding the Purpose of Data Cleansing

MLA Citations: "Data Cleaning: The Benefits and Steps to Creating and Using Clean Data." Tableau, www.tableau.com/learn/articles/what-is-data-cleaning.

Assessment:

As I continue my research into data visualization, I wanted to learn about data cleansing. The reason why I researched this subtopic of data visualization is because when I was talking with my mentor during our meeting, I had no clue what he was talking about. To better inform myself about data cleansing, I decided to read another article from the Tableau Software Company called, *Data Cleaning: The Benefits and Steps to Creating and Using Clean Data*.

Based on my previous research, I knew the purpose and benefits of visualizations and how to make an effective visualization. As I mentioned before, I have no clue what data cleansing was and how to use that for a data visualization. Luckily, this Tableau puts it all together.

In the first paragraph, the article highlights some of the key points from my previous research - such as the correlation between an effective data visualization and the development of a well-thought decision. Then, it

goes onto introduce the importance of data cleansing - also referred to as "data cleaning" and "data scrubbing" - and how this factors into the decision-making process (Tableau).

Next, the article provides a comprehensive understanding of what data cleansing is. According to the article, data cleansing is the "process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data" in a given set of data (Tableau). Even though there is a definition here, the article goes onto provide certain instances where data cleansing comes into play. The Tableau article highlights that when people combine different datasets for the purpose of creating their visualizations, there are certain things that the person creating the visualization should keep in mind. One such occurrence is the possibility of "duplicated or misleading" data - this has the potential to make the software/algorithms "unreliable, even though they may look correct" (Tableau). As the world of Big Data expands, the use of software is increasing too - so it is essential for a business to have accurate visualizations and that starts with the right data. Further in the article, it provides a template of sorts that ensures that data cleansing is done correctly and effectively. The reason why the article does not provide a clear-cut step-by-step guide on data cleansing is because the process of data cleansing will "vary from dataset to dataset" (Tableau).

In the process of data cleansing, there are five steps to ensure that the dataset that is being utilized for a visualization is considered clean. In step 1 of data cleansing, the dataset needs to be modified to "remove duplicate or irrelevant observations" (Tableau). This refers to the issue I mentioned above - the merging of different datasets (which can potentially create duplicate data). The occurrence of irrelevant data occurs when the "observations that do not fit into the specific problem" are being included in the analysis of the data (Tableau). The example that the article mentions is that of data involving the human generations - it highlights that if I was to collect data pertaining to millennials, but the dataset includes observations about the boomers or other generations. This step is essential to visualizations because it ensures that the data provided to the decision-maker(s) does not lead to them on the wrong path.

The next step to ensure a given dataset is considered clean focuses on the structure of the data itself as a potential issue. Step 2 of the data cleansing process is to "fix structural errors" (Tableau). Structural issues arise when the dataset contains problems in the categorical names associated with that dataset. These issues include: "strange naming conventions, typos, or incorrect capitalization" (Tableau). These structural errors should be corrected as the audience will have a better understanding of the data that is presented to them and they make their decisions.

As for the third step in this process, it aims to make the data as streamline as possible. In Step 3 of the data cleansing process, is to "filter unwanted outliers" (Tableau). This step is considered to be one of the most crucial steps because it has to deal with the problem that a person is attempting to solve with that very data. An outlier can be considered good or bad - it just depends on the situation. Removing an outlier can do two things: 1.) "help the performance of the data" or 2.) not "prove a theory" that is associated with the problem in question (Tableau). The choice of whether to remove the outlier is up to the individual and whether or not that outlier has a significant impact on the dataset.

The next step in this five-step process focuses on missing data. More specifically, Step 4 is to "handle missing data" (Tableau). The article emphasizes the importance of missing data and how it can mislead decision-makers to wrong decisions. So, there are some ways to manage missing data - but these ways are not considered optimal. The first way to manage missing data is to "drop observations that have missing values" (Tableau). However, removing such values from the dataset can potentially lead to misleading visualizations because of the reduction of information being presented. The other way that a person can manage missing data is to "input missing values based on other observations" (Tableau). However,

putting in values into the dataset can create another issue - bias - which can lead to a multitude of issues when it comes to the decision making process.

The fifth - and final - step of the data cleansing process is to validate the data. Validating the data in a given dataset is done through a series of questions that are used to check the accuracy of the data. Some of these questions include: “Does the data make sense?”; “Does the data follow the appropriate rules for its field?”; and “Does it prove or disprove your working theory, or bring any insight to light?” (Tableau).

As I move forward with my research into data visualization, I think I will need to talk with my mentor more about data visualization so that I can find out more about what I don’t know about the topic. As for right now, I think I need to look into data stories - which are another essential part of data visualization.