Understanding Educational Vulnerability in the Context of Disasters Using Visualizations

Cherish Caldwell

Follow this and additional works at: https://scholarworks.gsu.edu/iph_capstone

Recommended Citation
https://scholarworks.gsu.edu/iph_capstone/102

This Capstone Project is brought to you for free and open access by the School of Public Health at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Public Health Capstone Projects by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.
Abstract

Understanding Educational Vulnerability in the Context of Disasters Using Visualizations

By

Cherish Jocelyn Caldwell

August 23, 2018

BACKGROUND: Children are particularly vulnerable to the impact of natural disasters, yet limited scholarly attention has been placed on understanding their needs. The effect disasters may have on children’s educational attainment and achievement, otherwise known as educational vulnerability, is one of the least studied aspects of children’s disaster research. The use of visualizations using open access data repositories can facilitate researchers understanding of children’s educational vulnerability post-disaster.

AIMS: This paper illustrates how visuals can be used to address challenges that researchers may encounter when using educational datasets to evaluate disaster-related educational vulnerability. The challenges addressed include: (1) understanding data quality, (2) evaluating patterns within the data, (3) and evaluating for possible moderating variables.

DATA: This paper uses an example dataset containing educational data collected pre and post Hurricane Ike’s landfall in the Texas Gulf Coast in 2008. The publicly available data originated from the Texas Education Agency (TEA) and was compiled into a historical dataset for the school years 2003-2011. Schools served as the primary unit of analysis ($n = 464$). Performance on the Texas Assessment of Knowledge and Skills (TAKS) served as a proxy for school academic functioning.

CONCLUSIONS: The use of visualizations serves as a valuable method to aid in the understanding of educational vulnerability in the context of disasters. Visuals can be used to evaluate accuracy during data exploration, identify patterns within the data, and stimulate new questions and hypotheses. Future research should place focus on the utilization of longitudinal educational datasets, which will provide more detailed information regarding students’ educational vulnerability risks and needs.
Understanding Educational Vulnerability in the Context of Disasters Using Visualizations

by

Cherish Jocelyn Caldwell

B.S.N., University of Tennessee

A Capstone Submitted to the Graduate Faculty
of Georgia State University in Partial Fulfillment
of the
Requirements for the Degree

MASTERS OF PUBLIC HEALTH

ATLANTA, GEORGIA
30303
Understanding Educational Vulnerability in the Context of Disasters Using Visualizations

By

Cherish Jocelyn Caldwell

Approved:

Dr. Matthew Magee
Committee Chair

Dr. Betty Lai
Committee Member

Date: August 23, 2018
Acknowledgments

Dr. Matthew Magee, thank you for taking the time to be the committee chair on this capstone. Dr. Betty Lai, thank you for your guidance and support during my time at Georgia State. Thank you for always pushing me and challenging me to reach the next level of excellence. You have afforded me a multitude of opportunities for which I will always be grateful. You have been an extraordinary mentor.

I would also like to thank my family and friends for all of their support throughout this program. I want to especially thank my husband for his patience, understanding, and continuous support. I am truly appreciative of the sacrifices you have made during the last two years. Thank you to Hazel Shah for her assistance with understanding the grant on which this paper is based and for providing a theoretical foundation upon which I could build on. Thank you also to Philip Hudson for all of the feedback, support, and advice during the process of writing this capstone.
In presenting this capstone as a partial fulfillment of the requirements for an advanced degree from Georgia State University, I agree that the Library of the University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote from, to copy from, or to publish this capstone may be granted by the author or, in his/her absence, by the professor under whose direction it was written, or in his/her absence, by the Associate Dean, School of Public Health. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this capstone which involves potential financial gain will not be allowed without written permission of the author.

Cherish Jocelyn Caldwell

________________________________________
Signature of Author
Table of Contents

List of Figures .................................................................................................................. 7

Introduction ....................................................................................................................... 8

Example Data .................................................................................................................... 12

Methods ............................................................................................................................ 13

  Schools .......................................................................................................................... 13
  Procedures ...................................................................................................................... 13
  Measures ...................................................................................................................... 14

Visualizations .................................................................................................................. 16

  Challenge 1: Understanding Data Quality ................................................................. 17
  Challenge 2: Evaluating Patterns within the Data ..................................................... 20
  Challenge 2: Evaluating for Moderating Variables .................................................. 24

Conclusion ...................................................................................................................... 25

  Limitations .................................................................................................................. 28
  Future Directions ....................................................................................................... 29

References ...................................................................................................................... 31

Figures ............................................................................................................................ 39

Appendix A. R Code for Color Figures ........................................................................... 48

Appendix B. R Code for Black-and-White Figures ......................................................... 55
List of Figures

Figure 1A. Number of Students by School Year

Figure 1B. Number of Students by School Year for Campus #101910003

Figure 1C. Spaghetti Plot of TAKS Met Recommendation by School Type

Figure 1D. Spaghetti Plot of TAKS Met Recommendation by School Type (Truncated)

Figure 2A. Scatter Plot of TAKS Met Recommendation by School Year

Figure 2B. Jittered Scatter Plot of TAKS Met Recommendation by School Year

Figure 2C. Scatter Plot of TAKS Met Recommendation by Sociodemographic Variables

Figure 3A. Boxplot of TAKS Met Recommendation by School Type

Figure 3B. Boxplot of TAKS Met Recommendation by Test Subject
Understanding Educational Vulnerability in the Context of Disasters Using Visualizations

Natural disasters can have a significant impact on the health and well-being of children. Children are among the groups most vulnerable to the ramifications of disasters due to the developmental immaturity of their mental, social, and physical health (Peek & Stough, 2010; Weissbecker, Sephton, Martin, & Simpson, 2008). According to the United Nations Children's Fund (UNICEF), as many as 175 million children per year will be affected by disasters due to extreme weather events by the year 2023 (Burgess, 2013). This is an increase from the estimated 66.5 million children who were affected by disasters each year during the 1990’s. Despite the large number of children affected by disasters yearly, the needs and experiences of youth are often excluded from disaster research (Cox, Scannell, Heykoop, Tobin-Gurley, & Peek, 2017). Little scholarly attention has been placed on evaluating the effect disasters may have on children’s academic progress and educational outcomes; otherwise known as children’s educational vulnerability (Peek & Richardson, 2010). This paper serves as a first step in addressing disaster-related educational vulnerability using open access educational datasets. The primary focus of the paper is to discuss the use of visuals in exploring the data and deepening our understanding of children’s educational vulnerability. A sample dataset extracted from a publicly available educational data repository will be used to demonstrate the utility of visuals. For the purpose of this paper, educational vulnerability will be discussed solely in the context of disasters.

Educational vulnerability refers to “a number of negative post-disaster outcomes, such as missed school, poor academic performance, delayed progress, and failure to complete a program of study” (Peek & Richardson, 2010, p. S63). Destruction of school buildings and
student/teacher displacement can result in a loss of classroom time and an interruption in the mastery of key academic skills and concepts (Kousky, 2016; Lai et al., 2016). Furthermore, the absence of perceived security and stability through the loss of home or disruption of relationships with peers can be emotionally distressing (Dogan-Ates, 2010). This in turn can reduce long-term educational achievement and intellectual growth (Lai, Esnard, Lowe, & Peek, 2016).

There is a paucity of evidence-based literature regarding educational vulnerability in disaster research (Fothergill & Peek, 2015). This is surprising, as schools play a prominent role in the daily lives of youth, yet few studies have focused on educational attainment and progression post-disaster (Esnard, Lai, Wyczalkowski, Malmin, & Shah, submitted manuscript, 2018). Instead, the broader disaster literature tends to discuss schools primarily in the context of disaster preparation and strategic emergency response (Mutch, 2014). In addition, disaster vulnerability research has typically focused on the long-term physical and psychological damage inflicted as a result of the event (Peek & Richardson, 2010). Examples of such negative outcomes include illness, acute injury, posttraumatic stress disorder (PTSD), posttraumatic stress reactions, depression, and anxiety (Weissbecker et al., 2008; Mitchell & Bouchard, 2014; Pfefferbaum, Jacobs, Griffin, & Houston, 2015). If more studies acknowledged the impact of school disruption on children’s post-disaster recovery, more youth with educational vulnerability risks could be identified. Appropriate measures could then be taken to address negative post-disaster outcomes and reduce or eliminate academic decline (Pietro, 2018).

A key to addressing the knowledge gaps in the area of disasters and educational vulnerability is the ability to use educational data repositories. Measures commonly included in these datasets allow researchers to examine educational outcomes which may be affected by disasters such as academic achievement and attainment, school attendance and enrollment, and teacher staffing.
levels (Peek, 2008; United States Department of Education, n.d.). Changes in federal and state policies over the years have explicitly promoted the use of data, analysis, and research to improve educational systems. The increased demand and governmental financial support for educational transparency has resulted in an abundance of online data sources. This in turn provides more opportunity for educational vulnerabilities in disasters to be explored. According to Anderson (2005), increasing the availability of data-based evidence on children’s vulnerability outcomes results in a firmer basis for future practice and a more robust understanding of the societal impact of disasters in its entirety.

Despite the usefulness of educational datasets in understanding educational vulnerability, researchers and stakeholders often face challenges in understanding these large datasets. Using data visualizations may help to better understand educational data to address the challenges of educational vulnerability in the context of disasters. The National Forum on Education Statistics (NFES; 2016b) states that data visualizations “improve a data consumer’s ability to understand and analyze data, extract information, and use that information to make data-driven decisions” (p. 9). Data visuals take information that might otherwise be challenging to interpret or understand and convert it into a more easily comprehensible format (Ryan, 2016). The use of data visualizations is also important for effective communication and dissemination of results and newly discovered insights (Heer, Bostock, & Ogievetsky, 2010; Ryan, 2016). Before post-disaster educational vulnerability can be addressed, consumers must first understand the meaning of the data at hand. Involved stakeholders come from a variety of backgrounds, and it may be challenging for some to interpret the patterns, concepts, and results of this complex data (NFES, 2016b). The utilization of graphics and plots can take a wealth of information and transform it into a manageable and interpretable format for all stakeholders to understand.
Researchers may face three primary challenges when using educational datasets to examine educational vulnerability. The first challenge is understanding the quality of the available data. Visualizations help researchers explore whether or not a dataset is sufficient for the research question, and what changes need be applied to the data. Many systems contain databases which focus on specified levels of education (i.e., student, school, teacher, or fiscal) and may not be managed by the same department within an agency (Rudo, 2005). Inaccuracies, errors, or missing data may produce biased results. In addition, the data available may not perfectly capture the researcher’s question, and some information may need to be recategorized or excluded (Yorke, 2011). Exploring data through visualizations assists the researcher in inspecting the data, detecting dubious results, and confirming the appropriateness of the dataset for the research question at hand. Through visual exploration, researchers are more likely to discover revelations related to educational vulnerability which might otherwise go unseen (Smith, 2014).

Researchers may also have difficulty understanding the longitudinal patterns within an educational dataset. In order to fully understand the effects of disasters, it is most advantageous to examine risk factors and vulnerabilities both before and after the event (Masten & Osofsky, 2010). However, due to limited amounts of available data and the unpredictability of disasters, many studies can only examine outcomes over a short period of time after an event has already occurred (Kousky, 2016; Masten & Osofsky, 2010). The use of open access systems allows for the integration of data over weeks, months, or years, depending on the dataset (Chudagr, & Luschei, 2016; Daggett, 2014). This provides an opportunity for researchers to examine patterns and outcomes associated with educational vulnerability both before and after a disaster. It can be challenging for researchers to discern the patterns and relationships within a complex dataset by examining numbers alone (Francis, Jacobsen, & Friesen, 2014). However, visualizations can
enhance and highlight any change in pattern over time and make the results more meaningful (Chen, 2017).

The third challenge researchers may face is difficulty in understanding interaction effects within the data. Interaction occurs when an independent variable’s effect on a dependent variable differs according to the levels of a third variable, or moderating variable (Andersson, Cuervo-Cazurra, & Nielsen, 2014). Statistical testing must be used to verify interactions, but visuals can serve as an aid for researchers to examine potential moderating variables. This may be particularly useful for researchers who are unfamiliar with the subject matter or the statistical method used for analysis. Visualizations provide a fast and efficient method for detecting potential moderating effects before delving deeper into the data.

**Example Data**

The dataset used for this study was part of a larger study for a National Science Foundation grant, Award #1634234 (Principal Investigator Betty Lai, PhD). The publicly available data originated from the Texas Education Agency (TEA), the state agency which supervises primary and secondary Texas public schools (Texas Education Agency, 2018). The data was compiled into a historical dataset which contained data for the school years 2003-2011. The primary study aims were to examine the patterns and predictors of school academic recovery post natural disaster. Specifically, this study examined the effect of Hurricane Ike on academic performance, which made landfall during the 2008-2009 school year (Lai, Esnard, Savage, Shah, & Wyczalkowski, 2018). Standardized test scores, which served as a proxy for school recovery, and socioeconomic and physical vulnerability variables were extracted for school years pre and post-Hurricane Ike.
Method

Schools

Schools were used as the primary unit of analysis ($n=464$). The student population within these schools ranged from a minimum of 17 students to a maximum of 4,463 between the years of 2003 and 2011, the school years examined for this study. All grade levels between 3rd and 12th were included. Categories of schools were comprised of primary (1.08%, $n=5$), elementary (61.64%, $n=286$), intermediate (5.39%, $n=25$), middle (13.36%, $n=62$), junior high (4.96%, $n=23$), high (11.85%, $n=55$), K-12 (0.22%, $n=1$), and schools designated as other (1.51%, $n=7$) (Lai et al., 2018).

Procedure

Initial criteria for inclusion in the study was eligibility for the Hurricane Ike Provision, which resulted in an initial sample size of 623 campuses (Lai, et al., submitted manuscript, 2018; Texas Education Agency, 2009a). Due to the possible effect Hurricane Ike may have had on student academic performance, standardized scoring was adjusted for the 2008-2009 school year. The Hurricane Ike Provision allowed schools to receive a modified accountability rating if certain eligibility criteria were met. Eligibility for the provision included: a) schools located in one of 29 designated Hurricane Ike FEMA disaster areas, and b) schools which “were closed for ten or more instructional days between September 10, 2008 and late October 2008 (Texas Education Agency, 2009a). If the 2009 school rating was either lower than 2008 or deemed as Academically Unacceptable, then TEA would instead issue a rating of Not Rated: Other (Texas Education Agency, 2009a). Schools were then excluded from the study if they were not classified as a traditional public school, if there was a significant amount of missing data, if they
did not enroll children higher than 2nd grade, or if there were various irregularities regarding enrollment. Examples of enrollment issues included campuses which closed before or opened after Hurricane Ike or campuses which had an enrollment of zero from 2003 to 2011 (Lai et al., submitted manuscript, 2018). The final sample size after exclusions were applied resulted in 465 campuses and 55 districts.

**Measures**

**School academic functioning.** The Texas Assessment of Knowledge and Skills (TAKS), a standardized testing program administered to Texas public school students in the 3rd to 11th grade, served as a proxy for school academic functioning. During the spring semester of every school year, students were tested on several subjects including reading, mathematics, writing, science, English, and social studies (Texas Education Agency, 2011). This testing occurred during the years 2003 to 2011 (Lai et al., submitted manuscript, 2018). The specific subjects administered to students varied depending on grade level and the year of administration. Mathematics was the only test subject to be administered at every grade level (Texas Education Agency, 2011). As described by Lai et al. (submitted manuscript, 2018), reading was administered to 3rd-9th, writing to 4th and 7th, social studies to 8th and grades 10th-11th, and English/language arts to grades 10-11th. Science was administered to grades 5th and 10th-11th, although beginning in 2005 this subject was also administered yearly to 8th grade students.

**Scale score.** Student raw TAKS scores were converted to scale scores by the TEA. Score conversion allowed for the direct comparison of test results on a consistent scale among all students (Texas Education Agency, 2010). The following formula was used for scaling:
\[ SS_j = (\theta_j \times T1) + T2 \]

The Texas Education Agency (2010) described the estimate formula where “\(SS_j\) is the scale score for student \(j\), \(\theta_j\) is the Rasch partial credit model proficiency level estimate for student \(j\), and \(T1\) and \(T2\) are scale score transformation constants that establish the scale score system,” (p. 103).

**Accountability indicator.** The Accountability Indicator was the measure used to assess school performance by the TEA. The Accountability Indicator was obtained by calculating the percentage of TAKS tests out of all test taken which met specified test score cut off levels (Texas Education Agency, 2005). For the purpose of this dataset, two categories of scaled scoring cutoffs were used. *Met Standard* refers to the minimum TAKS passing scoring standard – a scaled score of 2100. *Commended Performance* refers to students who met a higher minimum standard - a scaled score of 2400 (Texas Education Agency, 2009b). Scores were summarized by test subject, test grade, or a summary of all combined scores by campus.

In 2002, the Texas State Board of Education adopted a new, more challenging performance standard. Scaled scoring which correlated to *Met Standard* was set at two standard error of measurement (SEM) below the new recommended minimum in 2003 and one SEM below in 2004 (Texas Education Agency, 2009b). The new standards were fully implemented in 2005. For 11th grade testing only, *Met Standard* was set at two SEM below the new recommendation for 2003 and 2004 and one SEM in 2005.

**Predictors.** Rates of potential risk factors which may affect academic performance were obtained for each individual school for the years 2003 to 2011.

*Limited English proficient (LEP).* LEP was the percent of students who had limited English proficient according to criteria set by the Texas Administration Code. These students
were identified by the Language Proficiency Assessment Committee (Texas Education Agency, 2011).

**Percent economically disadvantaged.** The economically disadvantaged variable was a TEA calculation of the percentage of students who were eligible for free or reduced-price lunch (Texas Education Agency, 2011).

**Percent minority.** The minority variable consisted of the percent of minority students by campus as calculated by study researchers. Minority students included African-American, Hispanic, Asian/Pacific Islander, or Native American (Lai et al., submitted manuscript, 2018).

**Percent mobility.** The mobility variable consisted of the percent of students who were members of the school for less than 83% of the school year. This corresponded to missing six or more weeks at a particular school (Texas Education Agency, 2011).

**Students with disciplinary placements.** Students with disciplinary placements included the percent of students placed in alternative education programs. This was calculated as a ratio of the number of students with one or more disciplinary placements by the number of students who were in attendance at any time during the school year (Texas Education Agency, 2011).

**Visualizations**

In the following section, visualizations were used to address the challenges that researchers may encounter when using educational datasets to evaluate educational vulnerability. The challenges described include understanding data quality, evaluating patterns within the data, and evaluating for possible moderating variables. In each section, open-source data from the Texas Educational Agency was used to illustrate how visualizations address these challenges by streamlining important information to build new knowledge. All plots and figures were created using the free open-source license program, R 3.4.3 (R Core Team, 2017). Data visuals were created using the R package, ggplot2 (Wickham, 2017). Figures were numbered to reflect the
challenge with which they were associated (e.g., Figure 1A illustrates Challenge 1, Figure 2A illustrates Challenge 2, etc.).

**Challenge 1: Understanding Data Quality**

Data quality can be defined as “data that [is] fit for use by data users or data consumers” (as cited in Wahyudi, Kuk, & Marjin, 2018, p. 458). Data quality assessment is essential in the creation of high-quality data which has both reliability and validity (Chen, Hailey, Wang, & Ping., 2014). High-quality data is important for instilling confidence in the researcher’s assessment of analysis outcomes and decision-making based on these outcomes. When using educational datasets to evaluate disaster vulnerability, the researcher may need to combine data from multiple sources. Each source of data may have varying levels of data quality or become compromised during the integration process. This highlights the importance of visual quality assessment to accurately evaluate educational vulnerability outcomes. Data quality can be classified into one of four categories: intrinsic, contextual, representational, and accessibility. In the next section, we will discuss how visuals can be used to evaluate two of the four data quality types: intrinsic and contextual.

**Visualizations to Address the Challenge**

**Intrinsic data quality.** Intrinsic data quality refers to inherent qualities of data such as accuracy, objectivity, and believability (Wayudi, Kuk, & Marjin, 2018). One way in which intrinsic data quality can be explored is by evaluating data for the presence of outliers. An outlier is a data point which lies a significant distance from other values (Sadik & Gruenwald, 2013). Outliers may be an accurate depiction of the data, or they may be created by human error, instrumental error, or a host of other factors. It is important to check for outliers when evaluating
data as they may lead to statistical bias or inaccurate results. In Figure 1A we plotted the number of students in each of the 464 schools included in the dataset from the year 2003 to 2010. The schools were colored according to school grade type. In the figure, one high school was attended by over 4,000 students, which was well above the number of students in the remaining schools. This school was considered an outlier. It is up to the researcher to determine if this outlier was accurate or created by error.

It was difficult to distinguish the outlier from the rest of the schools included in the dataset in Figure 1A. In order to better evaluate the outlier, visuals allowed the researcher to subset the data to only include schools attended by 4,000 or more students. Figure 1B shows the plot which was created from this subset dataset. There was only one school which met the subset criteria: campus identification number 101910003. Plotting the data point individually allowed the researcher to see a sharp increase in student population from the school year 2004 to the school year 2005. This led to further questions regarding the accuracy of this data. What may have caused an increase in student population after 2004?

**Contextual data quality.** Contextual data quality emphasizes the importance of evaluating data quality in accordance to the task at hand. In other words, contextual quality refers to the degree to which the data is applicable or helpful for the researcher (Wayudi, Kuk, & Marjin, 2018). The spaghetti plot created in Figure 1C portrayed the percentage of students who met TAKS recommended standards from the school years 2003 to 2010. The plot highlights the use of faceting to display groups of data. Faceting is a graphing function within ggplot2 in which subplots of discrete data are displayed side by side, allowing for ease of comparison (Chang, 2013). Figure 1C was facetted by school type with the mean percentage of students who met TAKS recommendation plotted for each year on the y-axis. The figure displayed a sharp
decrease in the average percentage of students who met recommendation in 2004. The decrease appeared consistent for all school types with a seemingly identical pattern throughout. The chances of such a drastic drop in scores for all school types during the same year seemed unlikely. This served as an alert to the researcher that the data warranted further investigation.

**Knowledge Gained from Visualizations**

The researcher took the information obtained from the visualizations displayed in Figures 1A and 1B to verify the data’s accuracy. Upon further investigation of the original data source, it was revealed that the outlying data point was indeed accurate. In the 2005-2006 school year two Texas high schools combined to become one large campus, identification number 101910003. This explained the sudden increase in student population. Through visualizations, the researcher was able to better evaluate and investigate outlying data and verify intrinsic data quality.

Upon further investigation of the pattern displayed in Figure 1C, it was discovered that there was a change in TAKS scoring from 2003 to 2005. In 2003, TAKS testing replaced the previous testing program, the Texas Assessment of Academic Skills (TAAS) (Texas Education Agency, 2009b). The TAKS examination was more challenging than the TAAS, so a transition plan was implemented to phase in the minimum *Met Standard* as schools became acclimated to the new program. This explained the apparent drop in test scores seen in Figure 1C. Although this would not be considered an error in the data, the results from these school years did not contribute to the evaluation of educational vulnerability. On the contrary, these results may have served as a visual distraction or a production of biased results. A solution to this problem was to truncate the data to begin with the school year 2005, as seen in Figure 1D. Through the use of visualizations, the researcher gained another tool to explore the data for accuracy and errors. The researcher
could then use the knowledge gained from visualization to make decisions regarding erroneous data or data that did not contribute to the study aims.

**Challenge 2: Evaluating Patterns Within the Data**

Pattern recognition is a key step in the evaluation of relationships among variables. The visualization of data makes it easier for researchers to distinguish these patterns among hundreds or thousands of data points and values (Ware, 2012). There are times when patterns within the data are not immediately apparent through the initial visual creation. When this occurs, interventions must be performed to provide an accurate evaluation of the data. The next section discusses ways in which this problem may be reduced or eliminated so that patterns may be explored. The section also describes ways to explore the patterns created by predictors within an educational dataset.

**Visualization to Address the Challenge**

**Pattern recognition.** In large datasets, pattern recognition is sometimes complicated due to problems such as overplotting. Overplotting occurs when scatterplot points obscure each other, impeding an accurate assessment of data distribution and pattern recognition (Chang, 2013). An example of overplotting is displayed in Figure 2A. Figure 2A is a scatterplot of students who met TAKS standards by year. Each individual point represents a single campus colored according to school grade type. The higher that the data point resided on the y-axis, the greater the percentage of students who met TAKS standards for that school year. Overlaid on the plot was a dashed horizontal line which represents the mean percentage of passing students for all school types combined. When visually assessing the mean trend from the 2003 to 2010, it was immediately apparent to the researcher that there was an abrupt change in pattern in 2004. As a result of the
visual exploration assessing data quality in Figure 1C, it was already known that the dip was due to a change in TAKS scoring that year. Following the visual assessment of the mean trend, the researcher could next evaluate for the detection of patterns among individual points within the scatterplot.

There appeared to be a positive trend in the percentage of students who met TAKS standards over the years. Viewing the plot from left to right, the lowermost data points, which represents the lowest performing schools, gradually began to increase in scoring from 2004 to 2010. At the top of the y-axis reside the highest performing campuses. It appears as though there was an abundance of data points at the top of the plot which are light green, the designated color for elementary schools. This was particularly true for schools in which 90% or more of students met TAKS standards. In contrast, very few non-elementary school campuses had greater than 80% of students who met TAKS standards. When evaluating for data accuracy in Figure 1C, the researcher was also able to visualize and compare the frequency of school types. When referring back to this figure, it appears that the sample included more elementary schools in comparison to other school types. This evaluation may have prompted the researcher to verify if Figure 2A encompassed an accurate depiction of the scatterplot pattern. Did elementary schools truly have a pattern of outperforming other school types or was the visualization obscured due to overplotting?

**Relationships among predictors.** The use of visuals is an effective way to explore the relationships between multiple predictors. In Figure 2C, a plot was created which displayed the percentage of students who met TAKS recommendation faceted by five student sociodemographic variables. The sociodemographic variables included students with disciplinary placements, economically disadvantaged students, students with limited English proficiency,
minority students, and students who have been school members for less than 83% of the year. The variables were labeled Conduct, Economic, Limited English, Minority, and Mobility respectively. Each facetted box displayed the percentage of students who met criteria for inclusion within the specified socioeconomic variable on the x-axis. The y-axis represents the percentage of all students who met TAKS standards for each socioeconomic variable. A color gradient was added wherein the darker purple represents the earlier school years with a gradual lightening of color for each subsequent school year (e.g., the darkest purple represents the year 2005 and the lightest year 2010).

**Knowledge Gained from Visualizations**

In order to evaluate the true pattern trajectory of the data, the researcher had to evaluate for the overplotting seen in Figure 2A. A solution to overplotting is the addition of random jitter to the individual data points (Chang 2013). Jittering adds a tiny bit of noise to each value to separate the overlaid points, as seen in Figure 2B. This figure showed that there were still many elementary schools which had a high percentage of passing students, but data points of other school types which were previously obscured were now seen. The plot also included the addition of loess curves for each school type, creating a more detailed and comprehensive visual. Primary schools were the only type to have a decrease in passing students over the years and elementary schools were indeed one of the top performing school types.

Visualization were also used by the researcher to examine the patterns and relationships between educational vulnerability risk factors and academic performance outcomes. The knowledge gained from these visualizations could be used as a basis for further statistical analysis. The following section describes the patterns visualized within Figure 2C:
Students with disciplinary placements. For the student disciplinary variable, it appears that there was more data available for the later school years in comparison to the earlier school years. There was no pattern seen for the earlier school years. Data from the later school years showed that as the percentage of students with disciplinary placements increased, the percentage of students who met TAK standards decreased.

Economically disadvantaged students. There was a pattern seen for the economically disadvantaged student variable. As the percentage of economically disadvantaged students increased, the percentage of students who met TAKS standards gradually decreased.

Students with limited English proficiency. When first exploring the data, one might have assumed that schools with a large percentage of students with limited English proficiency might have poor academic outcomes. The visualization of the data displayed results which were the opposite of what might have been expected. As the percentage of LEP students increased, the percentage of students who met TAKS standards also increased.

Minority students. As the percentage of minority students increased, the percentage of students who meet TAKS standards stayed steady until the percentage of minorities reached 75%. Once the percentage of minority students reached 75%, there was a sharp decrease in the percentage of students who met TAKS standards. This showed that minority percentage may not have had an impact on academic performance until the percentage reached a specific threshold.

Mobility status. As with the economically disadvantaged student variable, there was a pattern seen for the mobility student variable. As the percentage of mobile students increased, the percentage of students who met TAKS standards decreased sharply. This may be important when
assessing educational vulnerability after Hurricane Ike, as it is not uncommon for students to become displaced following a disaster.

**Challenge 3: Evaluating for moderating variables**

Evaluating for the presence of moderating variables is an essential step of data exploration. When looking at the effect of an independent variable on a dependent variable, this effect may change at different levels of a third, moderating variable. Research often evaluates the effect between the independent and dependent variables, but it is important for researchers to take moderating variables into consideration. Evaluating for moderation allows one to test the generalizability of results, identify the specificity of effects within subgroups, and to make decisions regarding interventions (MacKinnon, 2011).

**Visualizations to Address the Challenge**

Figures 3A and 3B were created in order to evaluate for the presence of moderation due to school type and test subject. Figure 3A displayed boxplots of the percentage of students who met TAKS recommendation for each school year. The axis for this plot was flipped with a vertical x-axis and a horizontal y-axis for better data visualization. The boxplots were then facetted by school type to see if there was a difference in TAKS scores at different levels of this third variable. On each boxplot, the dark red point represents the mean TAKS score. The various data points that lie beyond the boxplot minimum and maximum were considered outliers. As displayed in Figure 2A, there could be some overplotting of these outliers, so it was important for the researcher to take this into consideration when evaluating the visual. There was minimal spread within All (EE-12) and Primary schools. This may be due the fact that there were very few schools who fit into these categories as displayed in Figure 1C. After assessing the figure,
school type did appear to be a moderating variable. The means, medians, and overall distribution for each school type seemed to differ according to each group.

Next, Figure 3B evaluated the effect of testing subject as a moderating variable. The visual was created with the same x and y-axis as Figure 3A, with the exception of plots faceted by test subject instead of school type. Mathematics, science, reading, and social studies were the only subjects chosen for moderation visualization as writing and English were only administered to grades 4th and 7th and 10th and 11th respectively (Texas Education Agency, 2011). As with Figure 3A, this figure showed that there may have been a slight difference of means for test subject, but the difference was not as apparent. This may mean that test subject was not a moderating variable.

Knowledge Gained from Visualizations

Visuals may help the researcher identify potential moderating variables and to provide a better understanding of the moderating effects. Although visuals can aid in identifying potential moderators, the figures cannot be used to verify these effects. For example, by using only visualizations alone one could establish that there was a difference in variability and means between test subjects for each year of data in Figure 3B. However, when the researcher calculated the test subject means in R, all of the results for each subject were very close in number. The only way to confirm the significance of moderating effects is to run the appropriate statistical analysis method using the researchers chosen statistical software.

Conclusion

This paper provides an overview highlighting the various usages of visuals to aid in the understanding of educational vulnerability in the context of disasters. Educational vulnerability is a unique challenge in disaster research, as it poses a significant and particular risk to children
(Fothergill & Peek, 2015). Large quantities of educational data are continuously collected and stored by state agencies and school districts, providing a propitious opportunity for educational vulnerabilities to be explored (Alverson & Yamamoto, 2016, p. 3). Visualizations aid in the assessment of educational vulnerability outcomes by allowing researchers to evaluate data quality, patterns, and moderating variables which might otherwise go unobserved. The use of educational repositories served as an essential component of the visual exploration conducted in this paper. Using visuals based upon open source data increases the accessibility and transparency of analysis results in comparison to using complex statistical methods alone (Bendinelli & Marder, 2012).

It is beneficial to use visuals as a tool for data exploration throughout the entire analysis process, particularly when examining data quality as addressed in Challenge 1. It is commonplace for researchers to make use of visualizations solely to communicate analysis results. This may be due to time restraints or a lack of awareness regarding effective visualization techniques (Fox & Hendler, 2011). However, proactive visualization assessment can prevent the researcher from expending needless time and effort on data that cannot be used. For example, neglecting to check for data accuracy using a visual such as Figure 1C may have resulted in inaccuracies regarding our assessment of educational vulnerability. The evaluation of data quality is a crucial step of data exploration, as simple errors or unnecessary data can result in invalid statistical conclusions (Barchard & Verenikina, 2013). Through the use of visualizations, data quality was explored in a more comprehensive manner than if visuals had been used as an end-product of scientific analysis alone.

The identification of potential significant predictors, as seen in the evaluation of pattern recognition within the data in Challenge 2, is another benefit of visualization use. Economic
status, mobility, and minority status all appear to have an influence on the academic performance of TAKS test takers. Preexisting characteristics such as minority status, developmental disability, and low levels of social capital contribute to disaster vulnerability risk (Fothergill & Peek, 2015; Peek & Stough, 2010). This may result in a more difficult recovery for a child post-disaster, making it imperative that these risk factors are evaluated during data exploration. Visualizations can help to guide the selection of these predictors even before statistical model selection has begun (Ker, 2010).

As addressed in Challenge 3, visualizations can stimulate new questions and generate new hypotheses. This allows the researcher to become more familiar with the data and provides the ability to better interpret results (Ker, 2010). Data that may be used for research is abundant, yet researchers may grapple with the creation of hypotheses to extract knowledge from this information (Ma et al., 2017). The evaluation of moderating variables in Figures 3A and 3B serve as examples of visuals which may stimulate new analysis ideas for the researcher. If there is a difference in student academic performance based on school type, why might that be? Does student grade level affect long-term educational vulnerability outcomes? These questions may also be asked regarding testing subject. An additional example is the limited English proficiency visual seen in Figure 2C. As discussed in Challenge 2, an increasing percentage of LEP students within each campus resulted in a similar increase in the percentage of students who met TAKS standards. Does one’s country of origin or culture have a protective effect on academic performance and/or educational vulnerability outcomes? The researcher may formulate similar exploratory questions early in their analysis process due to the revelation of previously unknown visual patterns.
Limitations

Despite the push for data transparency, educational data is often aggregated by district, school, and class level due to student privacy concerns and a lack of available resources (Marsh, Pane, & Hamilton, 2006; National Center for Mental Health Promotion and Youth Violence Prevention, 2012; Jones & Southern, 2011). Aggregation refers to the summarization of individual data values for the purpose of statistical analysis (NFES, 2016a). For instance, our example dataset did not provide TAKS scoring according to student race. In its place the TEA included the percentage of students for each race who attended each individual campus or district. Aggregation makes it challenging to interpret how educational vulnerability may affect students on an individual level. This limits the ability to visually assess certain aspects of educational vulnerability in detail. Access to disaggregated educational data would be advantageous for exploring the effects of educational vulnerability on vulnerable subgroups and to identify opportunity gaps to which interventions may be targeted (National Commission on Asian American and Pacific Islander Research in Education, 2015).

Additionally, there is currently no required standard for the storage and reporting of educational data in the United States. In response to the United States Congress’ rejection of a national student level data system, the National Center for Education Statistics (NCES) created the Common Education Data Standards Initiative (CEDS) (Common Education Data Standards, 2018; Borden, Calderon, Fourie, Lepori, & Bonaccorsi, 2013) This initiative was a collaborative effort between local, state, and national organizations to develop common data standards for a key subset of educational variables within statewide longitudinal data systems (NFES, 2015). However, participation in this initiative is voluntary and it is the responsibility of the state or agency to determine how the standards are implemented (Data Quality Campaign, 2014). This
results in educational data systems with inconsistent state and local data standards and decreased comparability between systems. This can be especially detrimental when evaluating post-disaster outcomes as seen in the aftermath of Hurricane Katrina in August 2005 followed by Hurricane Rita in September 2005. The combined impact of these disasters caused the largest displacements of students in United States history (Dane, McCaffrey, Kalra, & Zhou, 2008). A lack of comparable student databases contributed to a delay of the resumption of school for some students which may have increased educational vulnerability (NFES, n.d.). Decreased comparability between systems also impedes the reproducibility of this paper’s examples using educational data from other states.

**Future Direction**

The dataset used in this paper utilized data from the 2003-2011 school years. The data contains only snapshots of summary statistics for each school year and grade level. More recently, there has been a push for states to create longitudinal databases which contain both student and staff level data. This data is collected across entities and is linked over time (Castro, New, & Wu, 2017; NFES, 2010). The use of these new, detailed systems as a basis for data visualization is the next step for the evaluation of educational vulnerability in the context of disasters.

Longitudinal databases may serve as a benefit to understanding educational vulnerability as children are often mobile following a disaster. Hurricanes Katrina and Rita serve as an example of the usefulness of longitudinal databases. The damage and destruction caused by these disasters lead to more than 196,000 students experiencing some form of school disruption (Dane et al., 2008). There were 22,000 students who were relocated to different districts in Louisiana and 45,000 who were relocated to other states such as Texas, Georgia, Florida, and Mississippi.
(Imberman, Kugler, & Sacerdote, 2012; Tian & Guan, 2015). Previous studies have evaluated risk factors and outcomes of public school students who reenrolled in the Louisiana school system following these disasters. Due to limitations in the data, these studies were unable to take into account students who moved to other states, dropped out of school, or attended private institutions. These students constituted more than a quarter of displaced students and may have been most at risk for poor educational vulnerability outcomes (Dane et al., 2008).

Although there will still be a disparity in the evaluation of students who drop out of school following a disaster, the use of longitudinal student level data would allow for the tracking of students across state lines and district. Individual student outcomes could then be tracked over years and across campuses regardless of location. Visualizations based on these longitudinal datasets can be used to better understand educational vulnerability risks and perhaps to visualize the statistical models created from these datasets. The end result will provide the opportunity for a more detailed evaluation of the relationship between predictors and student educational vulnerability outcomes, providing a significant contribution to disaster research.
References


[https://ceds.ed.gov/FAQ.aspx](https://ceds.ed.gov/FAQ.aspx)


[https://doi.org/10.1080/10824660802350169](https://doi.org/10.1080/10824660802350169)


[https://files.eric.ed.gov/fulltext/ED535227.pdf](https://files.eric.ed.gov/fulltext/ED535227.pdf)


[https://doi.org/10.1007/s11920-016-0683-4](https://doi.org/10.1007/s11920-016-0683-4)


Figure 1A.
Figure 1B.
Figure 1C.

TAKS Met Recommendation by School Type

Grades
- All (EE-12)
- Elementary
- High
- Intermediate
- Junior High
- Middle
- Other
- Primary

School Year

TAKS Met Recommendation (%)

04 06 08 10 04 06 08 10 04 06 08 10 04 06 08 10 04 06 08 10
Figure 1D.

`TAKS Met Recommendation by School Type`

- All (EE-12)
- Elementary
- High
- Intermediate
- Junior High
- Middle
- Other
- Primary

Grades:
- All (EE-12)
- Elementary
- High
- Intermediate
- Junior High
- Middle
- Other
- Primary

School Year:
- 05 06 07 08 09 10 05 06 07 08 09 10 05 06 07 08 09 10

TAKS Met Recommendation (%): 20 40 60 80 100
Figure 2A. 

TAKS Met Recommendation by School Type

Grades
- All (EE-12)
- Elementary
- High
- Intermediate
- Junior High
- Middle
- Other
- Primary

TAKS Met Standard (%)

School Year

2004  2006  2008  2010
Figure 2B.
Figure 2C.
Figure 3A.
Appendix A

Data Visualizations (Color)

Cherish Caldwell

```r
knitr::opts_chunk$set(echo = TRUE, warning=FALSE)
```

#knitr::opts is a global option that is applied to every chunk in the file
#echo controls whether or not code will be displayed in final document
#warning suppresses warnings from being displayed in final document

Resources:

- [http://www.cookbook-r.com/Graphs/](http://www.cookbook-r.com/Graphs/)
- [http://uc-r.github.io/ggplot](http://uc-r.github.io/ggplot)

Note: R is case-sensitive

Variables

- **CA311CA**: Students Who Met Commended TAKS Standards (%)
- **CA311RA**: Students Who Met Panel TAKS Standards (%)
- **CA311TA**: Students Who Met Accountability TAKS Standards (%)
- **campus**: Campus Identification Number
- **county**: County Identification Number
- **cpetallc**: Total Students
- **cpetdisp**: Students with Disciplinary Placement (%)
- **cpetecop**: Economically Disadvantaged Students (%)
- **cpetlepp**: Students with Limited English Proficiency (%)
- **cpemallp**: Mobile Students (%)
- **cpetminp**: Minority Students (%)
- **cpstexpa**: Average Teacher Experience (Years)
- **cpstkidr**: Number of Students Per Teacher
- **district**: District Identification Number
- **Grades**: School Type Based on Grade Levels
- **MATH_SUM**: Students Who Met TAKS Standards for Math (%)
• **READING_SUM**: Students Who Met TAKS Standards for Reading (%)
• **region**: Region Identification Number
• **sch_cat**: School Year (Factor)
• **sch_yr**: School Year (Numeric)
• **SCIENCE_SUM**: Students Who Met TAKS Standards for Science (%)
• **SOCIAL_SUM**: Students Who Met TAKS Standards for Social Studies (%)
• **testsub**: Testing Subject for Each School Year (Long Format)

Removing prior R Objects and Setting Working Directory

```r
rm(list=ls())
setwd("C:/Users/Cherish/Dropbox/Cherish Personal/Capstone")
```

Installing and Loading Packages

```r
#install.packages("haven")
library(haven) #For Reading SAS Dataset

#install.packages("ggplot2")
library(ggplot2) #For Generating Visualizations

#install.packages("reshape2")
library(reshape2) #For Data Shaping

#install.packages("dplyr")
library(dplyr) #For Data Manipulation
```

```r
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
## filter, lag

## The following objects are masked from 'package:base':
## intersect, setdiff, setequal, union
```

Reading in NSF Data from SAS Datasets

```r
#full_school<-read_sas("./final_3_19.sas7bdat")
#NSF<-read_sas("./m_plus_6_25.sas7bdat")
```

Permanently Saving Datasets as R Dataframes

```r
#save(NSF, file="./m_plus_6_25.rda")
#save(full_school, file = "./school.rda")
#save(testing, file="./testing.rda")
```

Loading Permanent R Dataframes Upon Program Restart

```r
load("./school.rda") #full_school
load("./m_plus_6_25.rda") #NSF
load("./testing.rda") #testing
```
Creating Color Vector to Be Applied to Plots

mycol <- c("indianred2", "olivedrab3", "lightseagreen", "mediumpurple2", "darkorange1", "royalblue1", "gold1", "hotpink3")

CHALLENGE 1: UNDERSTANDING DATA QUALITY

FIGURE 1A

a <- ggplot(NSF, aes(x = sch_yr, y = cpetallc, colour=Grades, group=campus)) + geom_line(aes(group = campus, colour=Grades), size = 1, lty=1) + geom_point(size=1, colour="darkorange3") + labs(title = "Number of Students By School Year", x = "School Year", y = "Students") + theme(plot.title = element_text(size=rel(1.5),lineheight=16, hjust = 0.5, face="bold.italic"), plot.subtitle = element_text(hjust = 0.5), axis.title.x=element_text(size=12), axis.title.y=element_text(size=12))+ scale_x_continuous(breaks=c(2003,2004,2005,2006,2007,2008,2009,2010))

#Adding color vector to plot
a + scale_colour_manual(values=c(mycol))

FIGURE 1B

#First we find the ID of the school with >4000 students
large_school <- subset(NSF, cpetallc>4000)

#Next we subset to include only the school with the ID that we looked up above
large_school <- subset(NSF, campus==101910003)

#Next, we will graph the number of students at this chosen campus over the years
lrg_plot<-ggplot(large_school, aes(x = sch_yr, y = cpetallc, colour="purple")) + geom_line(colour="mediumpurple3", size = 1.05, lty=1) + geom_point(size=3, colour="darkorange3") + labs(title = "Campus ID #101910003", x = "School Year", y = "Students") + theme(plot.title = element_text(size=rel(1.2),lineheight=12,hjust = 0.5, face="bold.italic"), plot.subtitle = element_text(hjust = 0.5))+ scale_x_continuous(breaks=c(2003,2004,2005,2006,2007,2008,2009,2010))

lrg_plot

FIGURE 1C

b <- ggplot(data = NSF, aes(x = sch_yr, y = CA311TA, group = campus, colour=Grades))
FIGURE 1D
#Subsetting to include only years 2005 and above
NSF_subset <- subset(NSF, sch_yr>="2005")

Trunc_yr <- ggplot(data = NSF_subset, aes(x = sch_yr, y = CA311TA, group = campus, colour=Grades))

Trunc_yr + geom_line() + stat_summary(aes(group = 1), geom = "point", fun.y = mean, shape = 18, size = 3.1, colour="hotpink3") + facet_grid(. ~ Grades)+ labs(x = "School Year", y="TAKS Met Recommendation (%)") +
ggtitle ("TAKS Met Recommendation by School Type") +
theme (plot.title =element_text(size=rel(1.3), lineheight=12,hjust=0.5, face="bold.italic", colour="black"), strip.background=element_rect(fill = "lavenderblush2"),
strip.text.x = element_text(face="bold" ,size=9)) + scale_x_continuous(
labels =c("05","06", "07", "08", "09", "10")) +
theme(axis.text.x = element_text(face="bold", size=8))+
scale_colour_manual(values=c(mycol))

CHALLENGE 2: EVALUATING PATTERNS WITHIN THE DATA

FIGURE 2A
c <- ggplot(data = NSF, aes(x = sch_yr, y = CA311TA, group = campus, colour=Grades)+geom_point()+ scale_colour_manual(values=c(mycol))+stat_summary(aes(group = 1), geom = "line", fun.y = mean,lty = 2.5, size = 1, colour="hotpink3")

C + labs(x="School Year", y="TAKS Met Standard (%)") + ggttitle ("TAKS Met Recommendation by School Type") +
theme (plot.title =element_text(size=rel(1.3), lineheight=12,hjust=0.5,"bold.italic", colour="black"), strip.background=element_rect(fill = "lavenderblush2"),
strip.text.x = element_text(face="bold" ,size=9)) + scale_x_continuous(
labels =c("05","06", "07", "08", "09", "10")) +
theme(axis.text.x = element_text(face="bold", size=8)+
scale_colour_manual(values=c(mycol))
FIGURE 2B

d <- ggplot(data = NSF, aes(x = sch_yr, y = CA311TA, group = Grades, colour=Grades))+geom_point(alpha = 0.35, position = position_jitter()) + stat_smooth(method = "loess", se=FALSE, size=1.1) + scale_colour_manual(values=c(mycol))

d + labs(x="School Year", y="TAKS Met Standard (%)") + ggtitle("TAKS Met Recommendation by School Type") +
theme(plot.title =element_text(size=rel(1.3), lineheight=12, hjust=0.5, face="bold.italic", colour="black"))+ scale_y_continuous(breaks=seq(0,100,10))

Creating Long Format Dataset for Figure 2C

student <- NSF_subset[, c("campus", "Grades","sch_yr", "cpetdisp", "cpetecop",
    "cpetlepp","cpetminp", "cpemallp", "CA311TA")]

student_long <- melt(student,id=c("campus", "Grades", "sch_yr", "CA311TA"), measure.vars = c("cpetdisp", "cpetecop",
    "cpetlepp","cpetminp","cpemallp"), variable.name="Student_Variable", value.name="Percentage")

#Renaming levels
levels(student_long$Student_Variable)[levels(student_long$Student_Variable)=="cpetdisp"]<-"Conduct"
levels(student_long$Student_Variable)[levels(student_long$Student_Variable)=="cpetecop"]<-"Economic"
levels(student_long$Student_Variable)[levels(student_long$Student_Variable)=="cpetlepp"]<-"Limited English"
levels(student_long$Student_Variable)[levels(student_long$Student_Variable)=="cpetminp"]<-"Minority"
levels(student_long$Student_Variable)[levels(student_long$Student_Variable)=="cpemallp"]<-"Mobility"

FIGURE 2C

e <-ggplot(student_long,aes(x=student_long$Percentage, y=CA311TA, group=campus, colour=sch_yr))+geom_point()

e + facet_grid(student_long$Student_Variable~., scales="free_y") +
labels(x="Students (%)", y="TAKS Met Standard (%)", colour="School\nYear") +
ggtitle ("TAKS Met Standard By Sociodemographic Variables") +
theme (plot.title =element_text(size=rel(1.3), lineheight=12, hjust=0.5,
face="bold.italic", colour="black"), strip.background=element_rect(fill="
CHALLENGE 3: EVALUATING FOR MODERATING VARIABLES

FIGURE 3A

```r
f <- ggplot(NSF, aes(x = factor(sch_yr), y = CA311TA, colour = Grades, group=sch_yr)) + geom_boxplot(outlier.size=1) + stat_summary(fun.y=mean, colour="darkred", geom="point", shape=18, size=1.5)

f + facet_grid(~Grades) + coord_flip()+ labs(y="TAKS Met Recommendation (%)", x="School Year") + theme(axis.text.x = element_text(size=8, face="bold"), strip.text.x = element_text(face="bold", size=9), strip.background=element_rect(fill="lavenderblush2")) + scale_colour_manual(values=c(mycol))
```

FORMATTING DATASET FOR FIGURE 3B

```r
#Subset to create smaller dataset
testing <- subset(full_school, select=c(testsub, MATH_SUM, SCIENCE_SUM, READING_SUM, SOCIAL_SUM, sch_yr))

#Subset data to remove NA's (missing data)
testing_sub <- subset(testing, !is.na(testsub))

testing_melt<-melt(testing_sub, id.vars = c("sch_yr"), measure.vars = c("MATH_SUM", "SCIENCE_SUM", "READING_SUM", "SOCIAL_SUM"), variable.name = "Subject", value.name = "CA311TA")

#Renaming levels for graph appearance
levels(testing_melt$Subject)[levels(testing_melt$Subject)=="MATH_SUM"]<="Math"
levels(testing_melt$Subject)[levels(testing_melt$Subject)=="READING_SUM"]<="Reading"
levels(testing_melt$Subject)[levels(testing_melt$Subject)=="SCIENCE_SUM"]<="Science"
levels(testing_melt$Subject)[levels(testing_melt$Subject)=="SOCIAL_SUM"]<="Social\nStudies"

FIGURE 3B
g <- ggplot(data = testing_melt, aes(x=sch_yr, y = CA311TA, colour=Subject))
```
g + geom_boxplot() + stat_summary(fun.y=mean, colour="darkred", geom="point", shape=18, size=1.5) + facet_grid(~Subject) + coord_flip() + labs(y="TAKS Met Recommendation (%)", x="School Year", colour="Subject") + theme(axis.text.x = element_text(size=8, face="bold"), strip.text.x = element_text(face="bold", size=10), strip.background=element_rect(fill="lavenderblush2")) + scale_colour_manual(values=c(mycol))
Appendix B

Data Visualizations (Black-and-White)

Cherish Caldwell

knitr::opts_chunk$set(echo = TRUE, warning=FALSE)

#knitr::opts is a global option that is applied to every chunk in the file
#echo controls whether or not code will be displayed in final document
#warning supresses warnings from being displayed in final document

Resources:
- [http://www.cookbook-r.com/Graphs/](http://www.cookbook-r.com/Graphs/)
- [http://uc-r.github.io/ggplot](http://uc-r.github.io/ggplot)

Note: R is case-sensitive

Variables
- **CA311CA**: Students Who Met Commended TAKS Standards (%)
- **CA311RA**: Students Who Met Panel TAKS Standards (%)
- **CA311TA**: Students Who Met Accountability TAKS Standards (%)
- **campus**: Campus Identification Number
- **county**: County Identification Number
- **cpetallc**: Total Students
- **cpetdisp**: Students with Disciplinary Placement (%)
- **cpetecop**: Economically Disadvantaged Students (%)
- **cpetelepp**: Students with Limited English Proficiency (%)
- **cpemallp**: Mobile Students (%)
- **cpetminp**: Minority Students (%)
- **cpstexpa**: Average Teacher Experience (Years)
- **cpstkidr**: Number of Students Per Teacher
- **district**: District Identification Number
- **Grades**: School Type Based on Grade Levels
- **MATH_SUM**: Students Who Met TAKS Standards for Math (%)
- **READING_SUM**: Students Who Met TAKS Standards for Reading (%)
• region: Region Identification Number
• sch_cat: School Year (Factor)
• sch_yr: School Year (Numeric)
• SCIENCE_SUM: Students Who Met TAKS Standards for Science (%)
• SOCIAL_SUM: Students Who Met TAKS Standards for Social Studies (%)
• testsub: Testing Subject for Each School Year (Long Format)

Removing prior R Objects and Setting Working Directory
```r
rm(list=ls())
setwd("C:/Users/Cherish/Dropbox/Cherish Personal/Capstone")
```

Installing and Loading Packages
```r
#install.packages("haven")
library(haven) #For Reading SAS Dataset

#install.packages("ggplot2")
library(ggplot2) #For Generating Visualizations

#install.packages("reshape2")
library(reshape2) #For Data Shaping

#install.packages("dplyr")
library(dplyr) #For Data Manipulation

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##     filter, lag
## The following objects are masked from 'package:base':
##     intersect, setdiff, setequal, union
```

Reading in NSF Data from SAS Datasets
```r
#full_school<-read_sas("./final_3_19.sas7bdat")
#NSF<-read_sas("./m_plus_6_25.sas7bdat")
```

Permanently Saving Datasets as R Dataframes
```r
#save(NSF, file="./m_plus_6_25.rda")
#save(full_school, file = "../school.rda")
#save(testing, file="./testing.rda")
```

Loading Permanent R Dataframes Upon Program Restart
```r
load("./school.rda") #full_school
load("./m_plus_6_25.rda") #NSF
load("./testing.rda") #testing
```
Creating Color Vector to Be Applied to Plots

bwcol <- c("#262C26", "#3E433E", "#565B56", "#6F726F", "#878A87", "#A0A1A0", "#B8B9B8", "#D1D1D1")

CHALLENGE 1: UNDERSTANDING DATA QUALITY

FIGURE 1A

a <- ggplot(NSF, aes(x = sch_yr, y = cpetallc, colour=Grades, group=campus)) +
  geom_line(aes(group = campus, colour=Grades), size = 1, lty=1) +
  geom_point(size=1, colour="gray23") +
  labs(title = "Number of Students By School Year", x = "School Year", y = "Students") +
  theme_bw() + theme(plot.title = element_text(size=rel(1.5), lineheight=16, hjust=0.5, face="bold.italic"),
                     plot.subtitle = element_text(hjust = 0.5), axis.title.x = element_text(size=12), axis.title.y = element_text(size=12)) +

a + scale_colour_grey(start = 0, end = .9)
# First we find the ID of the school with >4000 students
large_school <- subset(NSF, cpetallc>4000)

# Next we subset to include only the school with the ID that we looked up above
large_school <- subset(NSF, campus==101910003)

# Next, we will graph the number of students at this chosen campus over the years
lrg_plot<-ggplot(large_school, aes(x = sch_yr, y = cpetallc, colour="gray24") ) +
geom_line(colour="gray24", size = 1.05, lty=1) +
geom_point(size=3, colour="black") + theme_bw() +
labs(title = "Campus ID #101910003", x = "School Year", y = "Students") +
theme(plot.title = element_text(size=rel(1.2),lineheight=12,hjust = 0.5, fa
FIGURE 1C

b <- ggplot(data = NSF, aes(x = sch_yr, y = CA311TA, group = campus, colour=Grades)) + geom_line() + stat_summary(aes(group = 1), geom = "point", fun.y = mean, shape = 18, size = 3.1, colour="black") + facet_grid(. ~ Grades) + labs(x="School Year", y="TAKS Met Recommendation (%)") + theme_bw() + ggtitle("TAKS Met Recommendation by School Type") + theme(plot.title =element_text(size=rel(1.3), lineheight=12,hjust=0.5, face="bold.italic", colour="black"), strip.text.x = element_text(face="bold",size=9)) + scale_x_continuous(labels =c("03","04","06","08","10")) + theme(axis.text.x = element_text(face="bold", size=8))+ scale_colour_manual(values=c(bwcol))
FIGURE 1D

# Subsetting to include only years 2005 and above
NSF_subset <- subset(NSF, sch_yr="2005")

Trunc_yr <- ggplot(data = NSF_subset, aes(x = sch_yr, y = CA311TA, group = campus, colour=Grades))

Trunc_yr + geom_line() + stat_summary(aes(group = 1), geom = "point", fun.y = mean, shape = 18, size = 3.1, colour="black") + 
facet_grid(. ~ Grades) + labs(x="School Year", y="TAKS Met Recommendation (%)") + 
ggtitle("TAKS Met Recommendation by School Type") + theme_bw() + 
theme(plot.title =element_text(size=rel(1.3), lineheight=12,hjust=0.5, face="bold.italic", colour="black"), strip.text.x = element_text(face="bold",size=9)) + scale_x_continuous( 
  labels =c("05", "06", "07", "08", "09", "10") + 
  theme(axis.text.x = element_text(size=8, face="bold"))) + 
scale_colour_manual(values=c(bwcol))
CHALLENGE 2: EVALUATING PATTERNS WITHIN THE DATA

FIGURE 2A

```r
C <- ggplot(data = NSF, aes(x = sch_yr, y = CA311TA, group = campus, colour=Grades)) + geom_point() + scale_colour_manual(values=c(bwcol)) + stat_summary(aes(group = 1), geom = "line", fun.y = mean, lty = 2.5, size = 1, colour="black")

C + labs(x="School Year", y="TAKS Met Standard (%)") + ggtitle("TAKS Met Recommendation by School Type") + theme_bw() + theme(plot.title =element_text(size=rel(1.3), lineheight=12, hjust=0.5, face="bold.italic", colour="black") + scale_y_continuous(breaks=seq(0,100,10)))
```
FIGURE 2B

```r
D <- ggplot(data = NSF, aes(x = sch_yr, y = CA311TA, group = Grades, colour=Grades)) + geom_point(alpha = 0.35, position = position_jitter()) + stat_smooth(method = "loess", se=FALSE, size=1.1) + scale_colour_manual(values=c(bwcol))

D + labs(x="School Year", y="TAKS Met Standard (%)") + ggtitle("TAKS Met Recommendation by School Type") + theme_bw() + theme(plot.title =element_text(size=rel(1.3), lineheight=12, hjust=0.5, face="bold.italic", colour="black") + scale_y_continuous(breaks=seq(0,100,10))
```
Creating Long Format Dataset for Figure 2C

```r
student <- NSF_subset[, c("campus", "Grades","sch_yr", "cpetdisp", "cpetecop",
    "cpetlepp","cpetminp", "c pemallp", "CA311TA")]

student_long <- melt(student, id=c("campus", "Grades", "sch_yr", "CA311TA"),
    measure.vars = c("cpetdisp", "cpetecop",
    "cpetlepp","cpetminp","c pemallp"),
    variable.name="Student_Variable",
    value.name="Percentage")

#Renaming levels
levels(student_long$Student_Variable)[levels(student_long$Student_Variable)==
    "cpetdisp"]<"Conduct"
levels(student_long$Student_Variable)[levels(student_long$Student_Variable)==
    "cpetecop"]<"Economic"
levels(student_long$Student_Variable)[levels(student_long$Student_Variable)==
    "cpetlepp"]<"Limited English"
levels(student_long$Student_Variable)[levels(student_long$Student_Variable)==
    "cpetminp"]<"Minority"
levels(student_long$Student_Variable)[levels(student_long$Student_Variable)==
    "c pemallp"]<"Mobility"
```

**FIGURE 2C**

```r
e <- ggplot(student_long, aes(x=student_long$Percentage, y=CA311TA, group=campus, colour=sch_yr)) + geom_point()

e + facet_grid(student_long$Student_Variable~., scales="free_y") +
    labs(x="Students (%)", y="TAKS Met Standard (%)", colour="School\nYear") +
```
FIGURE 3A

```
f <- ggplot(NSF, aes(x = factor(sch_yr), y = CA311TA, colour = Grades, group=sch_yr)) + geom_boxplot(outlier.size=1) + stat_summary(fun.y=mean, colour="black", geom="point", shape=18, size=1.5)

f + facet_grid(~Grades) + coord_flip()+ labs(y="TAKS Met Recommendation (%)", x="School Year") + theme_bw() + theme(axis.text.x = element_text(size=8, face="bold"), strip.text.x = element_text(face="bold", size=9))+ scale_colour_manual(values=c(bwcol))
```
# Subset to create smaller dataset

testing <- subset(full_school, select=c(testsub, MATH_SUM, SCIENCE_SUM, READING_SUM, SOCIAL_SUM, sch_yr))

# Subset data to remove NA's (missing data)
testing_sub <- subset(testing, !is.na(testsub))

testing_melt <- melt(testing_sub, id.vars=c("sch_yr"), measure.vars=c("MATH_SUM", "SCIENCE_SUM", "READING_SUM", "SOCIAL_SUM"), variable.name="Subject", value.name="CA311TA")

# Renaming levels for graph appearance
levels(testing_melt$Subject)[levels(testing_melt$Subject) == "MATH_SUM"] <- "Math"
levels(testing_melt$Subject)[levels(testing_melt$Subject) == "READING_SUM"] <- "Reading"
levels(testing_melt$Subject)[levels(testing_melt$Subject) == "SCIENCE_SUM"] <- "Science"
levels(testing_melt$Subject)[levels(testing_melt$Subject) == "SOCIAL_SUM"] <- "Social Studies"

FIGURE 3B

g <- ggplot(data = testing_melt, aes(x = sch_yr, y = CA311TA, colour=Subject))

g + geom_boxplot() + stat_summary(fun.y=mean, colour="black", geom="point", shape=18, size=1.5) + facet_grid(~Subject) + coord_flip() + labs(y="TAKS Met Recommendation (%)", x="School Year", colour="Subject") + theme_bw() + theme(axis.text.x = element_text(size=8, face="bold"), strip.text.x = element_text(face = "bold", size=10)) + scale_colour_manual(values=c(bwcol))