**6-2 Assignment: Predictive Model Visualization**

Kayla Sacks

Southern New Hampshire University

DAT-430 Leverage Data for Org Results

Professor Aubrey Condor

October 13, 2024

## Introduction

Data visualization is an important part of data analysis as it makes it easier to understand the data we are working with and allows us to identify trends, patterns, and outliers quickly (Geeksforgeeks, 2021). There are many types of visualization that we use to show our data from line charts, scatter plots, and histograms to tree maps, heatmaps, and network graphs. What visual we will use will always depend on our data and what we are trying to show to our stakeholders that are not familiar with data analysis.

## Visualization Techniques

The importance of visualization techniques has become more apparent when studying data and data analytics as sometime statical information can be the same for different data sets but tell entirely different stories when visualized. When it comes to visualization tools that tell us about the importance of visualizing data, Anscombe’s quartet has, by far, had the greatest impact on me regarding why we can’t just trust the statistical information of a data set. Anscombe’s quartet consists of 4 data sets with the same mean, standard deviation, and regression line but, when the data points are plotted on a scatter plot, we can see a clear difference.

Figure 1 below shows the 4 data sets for Anscombe’s quartet along with the summary statistics. We can see how even though statistical observations are the same, the plots themselves are vastly different in terms of their distribution (Gupta, 2022).

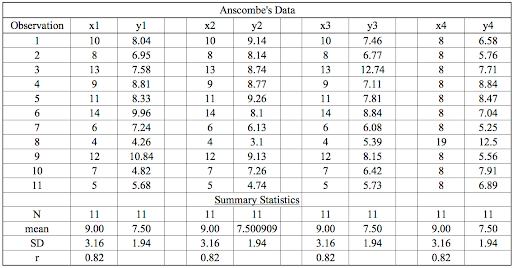


Figure 1 Anscombe’s quartet data sets with statistical summary <https://builtin.com/data-science/anscombes-quartet#:~:text=Anscombe%27s%20quartet%20is%20a%20group,you%20plot%20each%20data%20set>.

A group of graphs showing the results of a graph

Description automatically generated

Figure 2 Scatter plots of each data set with regression lines. <https://builtin.com/data-science/anscombes-quartet#:~:text=Anscombe%27s%20quartet%20is%20a%20group,you%20plot%20each%20data%20set>.

## Scenarios for Visualization

One of the most important situations to use Anscombe’s quartet is when explaining the importance of visualization to those new to data as the main point of the visualization is to express how a regression algorithm can be easily fooled and why visualizing the data is so important (Gupta, 2022). When trying to explain why visualizing data is important, this is a simple and easy visual that can be shown to anyone to help them understand quickly.

Another instance where this visualization would be effective before creating a regression model. Simply plotting each point on a scatter plot can give us an idea of the data so that we can easily see if a linear or non-linear relationship exists between the x and y. Plotting the points can also allow us to see outliers much easier. We may be able to spot the outlier in the fourth data set since both x and y values are so different then the rest, but it is harder to spot the outlier in the third data set just by looking at the table, this will become even more important as the number of observations, or data points, increases.

## Data Showcase

Anscombe’s quartet was created by Francis Anscombe to show how important it is to plot data before analyzing it and building your model (Gupta, 2022). Each data set in the quartet is graphed on a scatter plot (Figure 2) where we can see that each plot is clearly different even though the regression line is the same. The visualization allows us to see that the regression line does not fit the second or forth set at all. It also shows that the regression line fits the first set and shows an outlier in the third and fourth sets which would take much more time and effort to find if we just look at the data without the points being plotted. Without the visuals we would not understand the shape of the data.

## Creating Understanding

Anscombe’s quartet is one of the best ways to show stakeholders that are not familiar with data analysis why plotting data points and using visuals are so important in the analysis process and how visualizations can help us to better understand the data. One part that those not familiar with data analysis have trouble understanding is finding the patterns or trends of the data or why we need to use different data models. This visualization shows the stakeholders that even with the basic statistical information is the same, the story that the x and y points tell are completely different. This will allow stakeholders to make decisions based on a clearer understanding of the data since they can see the patterns visually instead of as numbers in a table or as statistics.

## References

Geeksforgeeks. (2021, July 1). *What is Data Visualization and Why is It Important?* GeeksforGeeks. <https://www.geeksforgeeks.org/data-visualization-and-its-importance/>

Gupta, S. (2022, February 9). *Anscombe’s Quartet: What Is It and Why Do We Need It? | Built In*. Builtin.com. <https://builtin.com/data-science/anscombes-quartet>