The Anscombe's quartet

The Answcombe's quartet named for the statistician Francis Anscombe was a string quartet active in England during the late 19th and early 20th centuries. The group was founded in 1882 by brothers William and Richard Anscombe and initially consisted of the two brothers and their cousin, Frank Anstey. The group later expanded to include Frank's brother, Ernest, and another cousin, George Anstey. The quartet was active until the early 1920's, when the Answcombe brothers retired. "The Anscombe's quartet is a great example of why it is important to always visualize your data before drawing any conclusions. The summary statistics for the four datasets are identical, yet the datasets look very different when graphed. This shows that summary statistics can be misleading and that one should always take a look at the data before drawing any conclusions." https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/anscombes-quartet-example/a/anscombes-quartet

It is a set of four datasets that have identical means, variances, and correlation coefficients, but have very different visualizations The summary statistics for Anscombe's quartet are: The mean of the x values is 9, mean of the y values is 7.5 variance of the x values is 11, the variance of the y values is 4, linear regression slope: 3.00, linear regression intercept: 0.500. The first dataset in Anscombe's quartet is a set of eleven x-y pairs. The second dataset is a set of eleven x-y pairs, all with x-values equal to 3. The third dataset is a set of eleven x-y pairs, all with y-values equal to 19. The fourth and final dataset is a set of nine x-y pairs, with x-values ranging from 0 to 2 and y-values ranging from 4 to 14. All four datasets have the same mean x-value (3), the same mean y-value (11.5), the same variance for x (11), the same variance for y (4.125), and the same correlation between x and y (0.816).

So, the first dataset talks of a linear relationship with low variance. The second dataset is clearly non-linear(quadratic), as the points are not evenly distributed along the line. The third dataset is again linear but has a much higher variance than the first dataset. The fourth dataset appears to be uniform. The fourth dataset has an outlier, which has a significant impact on the mean and variance. Anscombe's quartet demonstrates the importance of graphical data analysis The simple summary statistics give no indication of the differences between the datasets. The quartet also shows that outliers can have a significant impact on statistical properties. Anscombe's quartet is important because it highlights the importance of understanding the distribution of data before using traditional statistical methods on data that is not normally distributed. The methods can give incorrect results, and the results can be very different from one dataset to the next.

The four datasets are: The first dataset contains data that is clearly not normally distributed, the second dataset contains data that is less clearly not normally distributed, the third dataset contains data that is normal, and the fourth dataset contains data that is again not normally distributed. There are more robust statistical methods that can be used on data that is not Normally distributed, and Anscombe's quartet highlights the need for these methods. Some of the more commonly used methods include the bootstrap, the permutation test, and the Wilcoxon rank-sum test. Anscombe's quartet is also significant because it shows that even small changes in the data can have a big impact on the visualization. This is illustrated by the fourth dataset, which is very similar to the first but has a much different visualization.

The quartet also highlights the dangers of relying on summary statistics, such as the mean and standard deviation, to understand data. " Anscombe's quartet shows that summary statistics can be misleading and that one should always visualize the data before drawing any conclusions. "https://www.thoughtco.com/anscombes-quartet -3126545. We've discussed the

four datasets that were purposefully produced to illustrate the value of data visualization and how it may trick any regression algorithm. Hence, all the relevant features in the dataset must be visualized before putting any machine learning algorithm on them which will help to develop a decent fit model. This quartet, therefore, suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data, such as outliers, diversity of the data, linear separability of the data, etc. This informs us about the importance of visualizing the data before applying various algorithms to build models out of them.