Data Exploration and Analysis

PAPERS

Edris Safari | May 5, 2023

Table of Contents

Importance of Data Quality and what Constitutes Data Quality

It is widely known and accepted that Exploratory Data Analysis(EDA) is an essential part of any Data Science project. It is also equally natural to accept the fact that the quality of data is also of paramount importance to the success of the Data Science project.

Data extraction must always be followed by a data quality check. Data Source in Implementing Effective Data Quality defines the concept of "Data Quality Dimensions" as follows:

- **Completeness** : Remove corrupt, inaccurate, and impertinent data. For example customer addresses could be standardized in such a way so as to make queries such as which zip code, which city, which state, and even which street, county. Some components of the address may be missing(i.e. County, street name). We can deal with them appropriately without compromising the integrity of the data.
- **Conformity** : this is important when dealing with data from various regions, or geographic locations. For example shoe size in Europe , US, and UK are different. So some form of conformity needs to make sure that someone with shoe size of 42 in UK is not considered to have larger feet than a person with shoes size of 10 in the US.
- **Consistency**: Make sure that the values for an item are consistent in all data sources. For example credit rating in one system could be on a scale of 1 to 10 with 10 being excellent credit, and it could be A,B,C,D,E,F with F being terrible credit. The data quality analyst must decide how to reconcile this. Perhaps we could divide 10 by 6(number of letters in the letter system), and decide that those with credit rating 0 to 1.66(10/6) are assigned letter 'F', and those with credit rating 1.66 to 3.33 get 'D', and those with 3.33 to 4.99, get 'E', and so on. This would probably make the credit rating a bit more stringent, but other considerations are on the analyst's plate.
- **Accuracy**: incorrect data results in incorrect assumptions and also inaccurate models. It could also result in data being thrown away because it doesn't conform to a standard. For example if expected address is street number, followed by street name(i.e. 123 Lucas Blvd.) and someone's address happens to be Spring Lane(no street number), the 2nd person's address is deemed invalid and perhaps ignored in the overall analysis. Imagine there be 100's of thousand addresses that don't conform.
- **Duplicates**: This happens when there is duplicate value for the same set of attributes. For example, in the table below the customer has had two duplicate records show up in the report. This could be due to the SQL query and the tables included in the SQL. The analysts need to revisit the SQL and decide what to do. Just removing the duplicates may not be the appropriate approach.

order_id	customer_name	part_number
1234	John Smith	P1
1234	John Smith	P2
1234	John Smith	P3

1234	John Smith	P2

- **Integrity**: Integrity is reliability. If we cannot rely on data, then we as well might just pack it up and go home(and rethink our strategy-NOT GIVE UP). Data integrity is measured by its completeness. It is complete when well defined and agreed upon standards are adopted. Data integrity and the maintenance thereof, ensures that the data that the data analysists deal with are unaltered from the time they are created to the time they are received. Missing data is a simple example. Let's say there is a table that stores data by transactionID as primary index and another table has transactionID and customerID as indices. Merging these tables results in a report showing customerID, and transactionID, but some records are missing customerID. This data cannot be relied on. We must revisit data storage and retrieval procedures(while drinking a lot of coffee!)

Poor Data Quality leads to Poor Decision

Data quality is not the quality of the data per se. It is the quality of using or viewing the data and making correct conclusions and asking the right questions. It is leaving little or no room for errors.

One of the primary examples that comes up in the literature is the 'Simpson Paradox'-this is Simpson in the TV show "The Simpsons". It is typically illustrated by a law suit that University of Berkley averted because they proved the quality of data that showed they were guilty of discrimination in their admissions of women candidates was inadequate. A closer look at the data was a relief to the university officials who were afraid of getting sued if they went by the report below from their Fall of 1973 semester:

Sex	Applicants	Admitted
Men	8442	44%
Women	4321	35%

It showed 44% of men were admitted to only 35% women. University of Berkley asked Peter Bickel who is now a professor emeritus of statistics at Berkeley to analyze the data and find out which department was the culprit(a good question wouldn't you say?). They gathered their data and came up with the table below:

Department	# of Men	# of Women	Men Accepted	Women Accepted
A	825	108	62%	82%
B	560	25	63%	68%
C	325	593	37%	34%
D	417	375	33%	35%
E	191	393	28%	24%
F	373	341	6%	7%

Total	8442	4321		

Departments A,B,D, and F admitted more women than men. What, a discrimination against men? Not so fast. A closer look at departments A and F shows that department A has a higher acceptance rate than department F. However, the percentage of women(108/4321=2%) is much less than men's(825/8442=10%). Looking at department F in the same way shows that even though the percentage gap between women and men applicants is close(8% and 4% respectively), the acceptance percentages of 6% and 7%are dismal.

The percentage of women who applied to department F is higher than men(as opposed to those in department A). So the revelation is that a larger portion of women applied to a low-accepting department and lower portion of them applied to higher-accepting department. This is the cause of the misinterpretation of the data-Simpson's Paradox.

This sort of phenomena can affect any statistical analysis exercise and aversion of it is detrimental to the success of the project. In 2014, GM sent out recall notices to owners of cars whose defects had caused 13 fatal accidents. The problem was that they also sent these notices to some of the families of victims in those fatal accidents. A family had even moved to a different state, but they still received not one, but three recall notices(one for each defect). In another case Vodafone sent out erroneous bills to customers who had used their service abroad(perhaps a problem with Data Conformity or accuracy?).

Dealing with data in the data science realm is a responsibility that increases proportionally with respect to its affect(or influence) on the public . Sound ethical as well as computational considerations lead to quality and quality is the ever evasive holy grail that quality seekers seek. Luckily, in data science we have statistics and behind it, we have the computational power of the computer.

References:

Simpson's Paradox and Statistical Urban Legends: Gender Bias at Berkeley

From <https://www.refsmmat.com/posts/2016-05-08-simpsons-paradox-berkeley.html>

How UC Berkeley Almost Got Sued For SEX Discrimination….LYING Data?

From <https://medium.com/@dexter.shawn/how-uc-berkeley-almost-got-sued-because-of-lying-data-aaa5d641f571>

Kelleher, John D.. Data Science (MIT Press Essential Knowledge series) (p. 93). The MIT Press. Kindle Edition.
Data Source-Implementing Effective Data Quality

What is Data Integrity? Learn How to Ensure Database Data Integrity via Checks, Tests, & Best Practices <https://www.veracode.com/blog/2012/05/what-is-data-integrity>

Data Source - Implementing Effective Data Quality

GM apologizes for sending recall notices to victims' families-
https://www.autonews.com/article/20140603/OEM11/140609934/gm-apologizes-for-sending-recall-notices-to-victims-families

Vodafone customers wrongly told they owe £1,000s after using phones abroad-
https://www.moneysavingexpert.com/news/2019/10/vodafone-customers-incorrectly-told-they-owe-p1-000s-after-roami/

Distributions and Descriptive statistics

The statistical approach to solving data science problems is in effect looking at a problem, any problem, anecdotal or not from the perspective of a statistician. It is to model the real-world problem in a way that we can perform calculations. This is what gave rise to mathematics which is a foundation of all sciences dealing with our everyday(and future) lives.

We need data to perform the necessary calculations. Without data, we are left with a bunch of equations starring at us from the terminal screen or paper notebook. Allen Downey lists 5 tools to use when we take a statistical approach. These steps are in line with the CRISP-DM model. They are:

1. **Data collection** - We need data. There is always data, but not clean data, but it is data nonetheless.
2. **Descriptive statistics** - This is the visualization of data. Using a sample of the data, we create histograms, bar charts, box plots or other visualization techniques.
3. **Exploratory data analysis** - Hand-in-hand with descriptive statistics, we look for patterns that lead us to answer the original question , refine the question, come up with new ones, or change of plan and/or data source.
4. **Estimation** - once we establish the descriptive statistics, we take a sample of the data set(say 20%) to estimate the characteristics of the entire dataset(or population)
5. **Hypothesis testing** - This is when we test the hypothesis that we have developed in the last steps. We run the model(say linear regression or logistic regression, etc.) against the population data set(not sample) and evaluate differences between runs. This is when we look at statistical parameters such as p-value, adjusted R-Squared, false positives, false negatives, and other statistical values.

Descriptive statistics and inferential statistics are part of the statistical analysis that gives us confidence that our hypothesis testing will succeed. This is a repetitive process of taking a sample of the data from the population(master dataset), analyze and visualize them to make sure the sample and population agree. For example if we know that in a data set 40% are say in favor of issue A and 50% are against. If we take a sample of the population and the statistics shows a 20% to 80% ratio, we must look at a different sample or rethink our strategy.

The statistical analysis starts with distributions in the dataset. A distribution describes all possible values for a set of data and how often those values occur. For example if we toss a die 6 times, and the number 2 occurs one time, 3 occurs 2 times, and 4 occurs three times is a distribution. The number of children born to age groups of teens, 20's,30's,40's, and 50's is also a distribution. We can graph a distribution and visualize where if any data is concentrated and where the anomalies are. If we can get what is called a normal distribution, we can then do some estimation and test out hypothesis.

Normal distribution is a bell curve that shows the spread of data around the mean of the population. It shows some sample drifting away from the mean , but most will fall closer to the mean. The statistical summary that goes into creating a normal distribution determined by

Measures of Center and of spread. They are the statistical summary data that analysts wrestle will on a daily basis(from what I've read!). A brief description of each is given below:

Measures of Center:

Mode: The data value that is most frequently observed in the dataset. For example in the distribution of children born case above(Figure 2-3 in Think Stats: Exploratory Data Analysis) the mode would fall in the 20 to 28 year old giving most number of births.

Median: The median is the value of the middle position or index in the ordered list of dataset. For example if we have a set of 9 numbers ordered from lowest to highest in this way

(123,134,154,155,**158**,160 ,162,162,165)

Then Median = (10+1)/2=5, so the 5th element in the list is the median- i. If the number of values is even, the median index is X.5, so we take the average of the two values from index (X.5 - .5) and (X.5+.5)

Mean: Mean is the sum of all values in the dataset divided by their number. This is an average, but Downey make a good distinction. The analogy of the pumpkin example makes sense. In such a case, perhaps its more prudent to take the median as the center point. This way the 520 lbs. pumpkin would show as an outlier and can be removed from the calculations.

Measures of spread

Range: Range the difference between the max value and minimum value. This value by itself doesn't reveal anything because the range could be wide for cases such as the pumpkin example. This is where variance and standard deviation come in.

Variance: Variance is the sum of the squares of the difference between the values and the mean divided by the number of samples. This value shows the spread of data or mean squared deviation of each data point from the center.

$$S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

Standard Deviation: This S or the square root of the variance. It show how close are the values to the mean(center). The higher the standard deviation, the more wide spread the data are and the lower, the more compact.

With summary statistics, and the accompanying graphs, we can make inferences. This is the process of drawing conclusions about population parameters based on a sample taken from the population. The accuracy of the summary statistics is judged by how close the sample statistics are to the population parameters-that is their respective mean and standard deviation.

I found this topic to be very challenging, but a necessary knowledge to pursue. Learning to program these concepts would definitely help with the learning curve. Luckily, I found a lot resource that proved valuable information.

CRISP-DM Help Overview- https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm

Mode, Median, Mean, Range, and Standard Deviation (1.3) - Mode, Median, Mean, Range, and Standard Deviation (1.3)

Understanding Hypothesis testing, p-value,test-statistics help-Understanding Hypothesis testing, p-value, t-test - Statistics Help

The Shape of Data: Distributions: Crash Course Statistics - The Shape of Data: Distributions: Crash Course Statistics #7

Downey, Allen B.. Think Stats: Exploratory Data Analysis . O'Reilly Media. Kindle Edition.

Distributions and Distribution Functions

We've learned that a distribution is a set of finite observations (i.e. head or tail) that occur a number of times after a finite number of observations are made (i.e. tossing a coin 10 times). We've used histograms, simplified them with PMFs and CDFs to make better sense of the data by looking at its shapes. With PDM, we calculated the probability of an observation rather than the number of observations. The value of the probability being between 0 and 1 gave us a better view of the distribution. We evaluated mean, and variance of the PMF whose functions are bit different than the distributions that we can plot histograms for. The mean of a PMF distribution is

$$Mean(x) = \sum PMF(xi) * xi$$

, and the variance is

$$S^2 = \sum PMF(xi) * (xi - mean(xi)$$

While PMF works well for distribution with a small sample space, it does not for distributions with non-discrete and thus larger sample space. Cumulative Distribution Functions (CDFs) can be used to mitigate this issue. The CDF is a function that maps a value to its percentile rank. Percentile rank is a fraction of scores or values less than or equal to a given value. The equation for percentile rank is

Percentile_rank = 100*Count/number of values in sample set

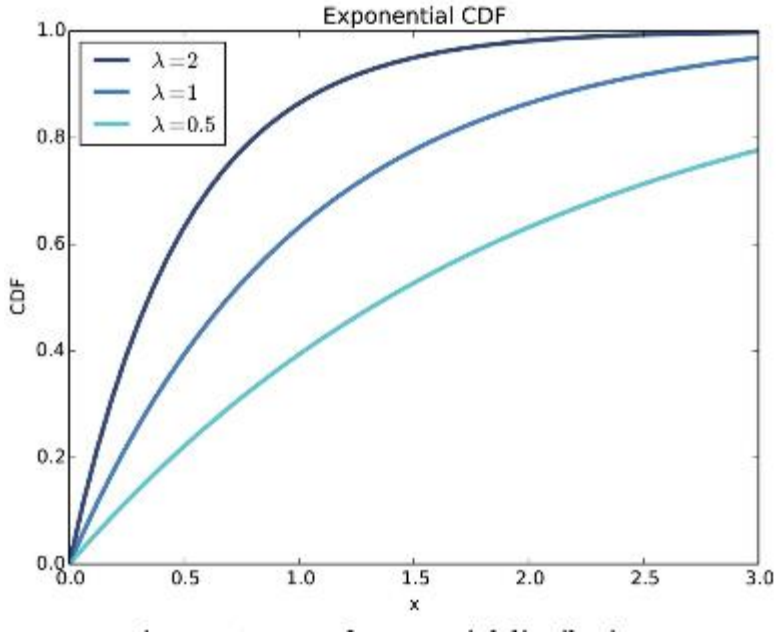
Count is the number of values in the distribution that are less than or equal to a given value. So if x is in the 90th percentile rank, x is greater than equal to 90% of the values in the distribution. The CDF of a value x in a distribution is percentile_rank/100 making the value of CDF between 0 and 1-as in PMF but with larger sample sets.

The distributions in PMF and CDF are based on empirical or verifiable distributions with finite samples. When faced with larger distributions and infinite samples, we resort to computing the properties of the distribution using a CDF function. As we know, CDF is the probability that an outcome is less than or equal to a given value.

When a distribution is composed of a series of events and there are measured times between events, then that distribution is called an exponential distribution. For example the amount of time it takes for a car to pass through a tollgate or the amount of time it takes before the next customer walks in or visits a web site have an exponential distribution. The CDF function for exponential distribution is:

$$CDF(x) = 1 - e^{-\lambda x}$$

Where the parameter
λ is the mean of the distribution and it value shapes the distribution as shown below:



Exponential distribution is the inverse of Poisson distribution which has the property of the number of events in a given time period(i.e. number of cars passing a tollgate in an hour). Usually we have a PMF and/or a CDF distribution for the Poisson distribution with a given mean value called lambda. For example, mean of 3 cars passing a tollgate in one hour.

Notice that the values in the Poisson distribution are discrete. To make them continuous so we consider the amount of time before the occurrence of an event and use the reciprocal of the mean of PMF λ . So instead of saying 3 cars per hour, we can say 20 minutes for 3 cars or $\mu = \frac{1}{\lambda}$. This turns the equation above to:

$$CDF(x) = 1 - e^{-x/\mu}$$

The resulting exponential distribution allows us to calculate how many cars within 25 minutes or after 12 minutes. However, we cannot calculate the number of events at an exact time such as at 15 minutes.

One usage of exponential distribution is in reliability. For example, in a factory the machines that perform various tasks have parts that need regular maintenance. Parameters such a mean time before failure and mean time to repair contribute to the overall equipment effectivity (OEE). With these measures, maintenance can be predicted and scheduled rather than reacting to a failure which is costly in terms of equipment down time.

References

The Difference Between Poisson and Exponential Distributions From <https://www.youtube.com/watch?v=Z-8FtjZNlb4>
The Exponential Distribution From https://courses.lumenlearning.com/introstats1/chapter/the-exponential-distribution/
Probability Exponential Distribution Problems From <https://www.youtube.com/watch?v=J3KSjZFVbis>
Exponential Distribution! Definition | Calculations | Why is it called "Exponential"? From Exponential Distribution! Definition | Calculations | Why is it called "Exponential"?

Scatter Plots to visualize multiple variables

With single variables, we use histograms, PDMs and CDF to visualize the distribution of data. With multiple variables, visualizing the data in scatter plots is ubiquitous in exploratory data analysis. Scatter plots show three characteristics that we can use to understand the relationship between two variables. The two variables in question have values which are represented by records in a dataset. The dataset can be of a whole population or a subset of it(called sample dataset). Some variables are what are commonly called predictors or independent variables. These independent variables can affect the behavior of another variable in the dataset. That variable is called the outcome or the dependent variable. In a one independent variable and one dependent variable situation, the equation that can best be used to discern any intuition from the dataset it the simple linear equation.

$$Y = mX + b$$

Where X and Y are the independent and dependent variables respectively and the m and b are the slope and intersection of the line drawn in the X & Y axis. This equation is also written in the following form:

$$Y = b0 + b1X + Noise$$

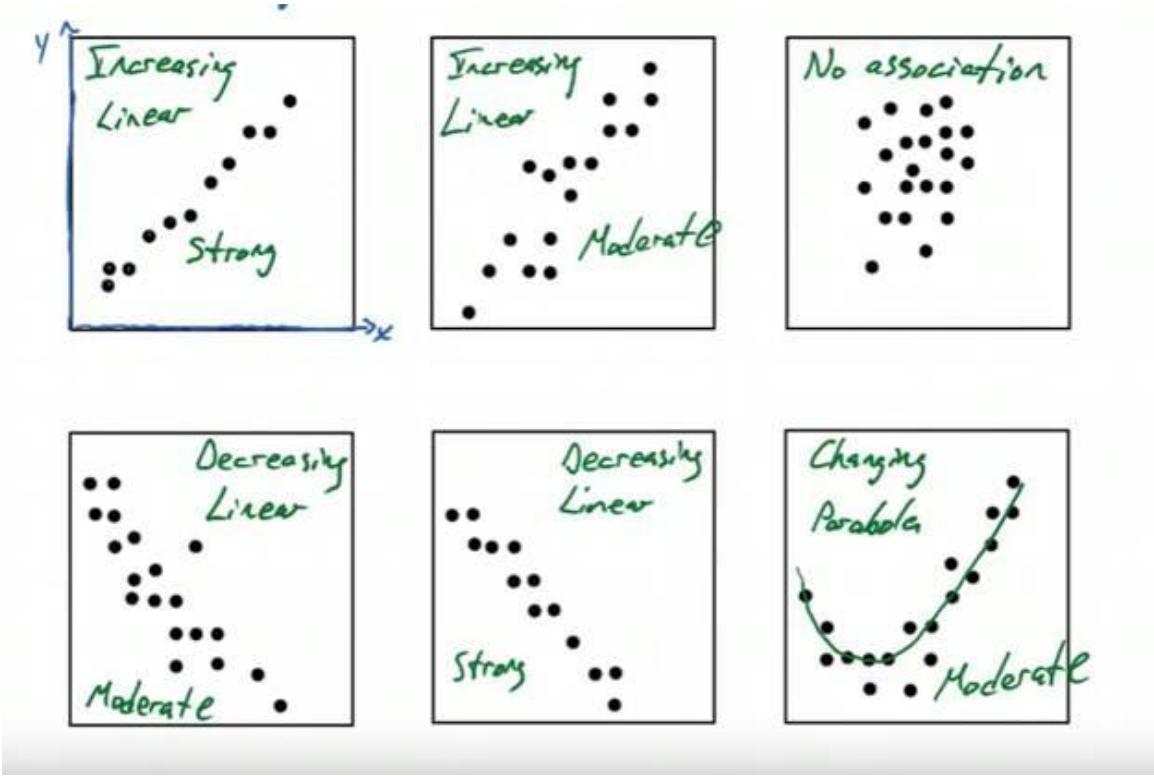
Where the slope and intersection are now called the b values and noise is to take care of things that give variability to the equation. Things such as measurement whose accuracy is dependent on something like heat or even the respondent's state of mind or understanding the question. The b values are used in determining correlation and a host of other information from the graph. In case of multiple independent variables, the equations become:

$$Y = b0 + b1X1 + b2X2 + \ldots + bnXn + Noise$$

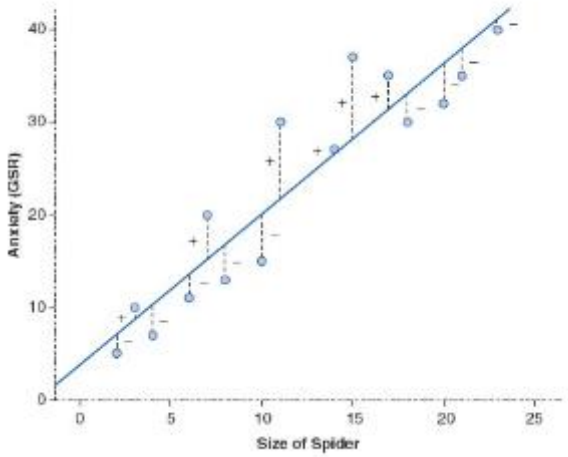
So, where do scatter plots come in? The answer is at the very beginning of the data exploration phase. Scatter plots show association between two variables using three visual tools:

1. Trend
2. Shape
3. Strength

The picture below (Ref1) shows the three characteristics. In an ideal situation, we want the trend to be increasing or decreasing, the shape to be somewhat linear (follow a straight line), and the strength to be strong or moderately strong. The strength is shown by the number of points that are closer to the "imaginary" line. As shown below, there are cases where the scatter plot shows no association or linearity. In the case where no association is shown, we can still try to decipher something by say binning. Binning is when we can group sets of data within a range. For example, grouping respondents in their age group of 20 to 30, 30 to 40, 40 to 50, and so on.



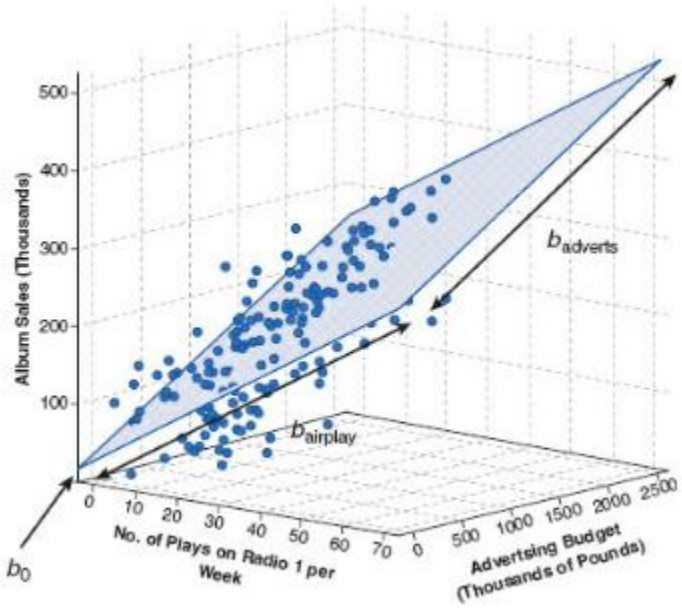
When we view scatter plots and draw that imaginary line to show trend, shape and strength, we automatically see distances between the points in the chart and the imaginary line. This is where the concepts of deviation, and variance come to the picture. From there we can calculate standard deviation which shows how far a spread the actual points are from the imaginary line. Picture below (Ref 3) shows the variance of actual data and the line that best fits the plot.



In some cases, the scatter plot can be adjusted by jittering. Jittering is when we suspect that the data is not fully representative of the variable. For example, if we asked for height in full inches rather than including fraction of inches or weight in full pounds rather than pounds and ounces, we would have introduced noise in the data which needs to be compensated. Jittering adds random noise to reverse these effects.

Sometimes jittering causes saturation. Saturation happens when the data is too dense in some areas and jittering hides data that could otherwise be useful. Adjusting the plot's transparency by using alpha setting in the plot, we can see the overlaps where the plot Is darker. This could be effective is smaller datasets, but in larger datasets the plot would not be effective,

Scatter plots are effective tools to visualize the data and make assumptions and hypothesis. They are by no means effective is determining correlation between variables. It becomes even more complex when we are dealing with multiple independent variables as shown in picture below(ref3). Here, we have two independent variables (adverts and airplay) that predict/determine Album Sales. While the scatter plot shows nice trends, shapes and strengths, the ultimate correlational analysis lies in concepts such as correlation, covariance and the two main statistic Pearson correlation for linear relationships and Spearman's Rank correlation for non-linear relationships. These are topics for other discussions which are a continuation of what starts at scatter plots and their derivatives.



Ref1 - Statistics Scatter Plots & Correlations Part 1 - Scatter Plots
Ref2- Downey, Allen B.. Think Stats: Exploratory Data Analysis . O'Reilly Media. Kindle Edition.
Ref3- Field, Andy. Discovering Statistics Using R (p. 255). SAGE Publications. Kindle Edition.