CHAPTER – 4
EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a critical step in the Analytics process. This would be the first formal output and can act as a mirror giving feedback to the business. Although business managers are on top of the facts about the business, EDA always bring out facts that will surprise them. In general, the EDA involve analysis of data that will answer the question 'what is going on?'

EDA relies heavily on visual presentation of results as discovery of patterns is a major objective at this stage. We look for evidences in the data so that a tentative understanding is developed about what influences the phenomenon we are examining and how it is influenced. As part of the analysis objective, we will conduct formal hypothesis tests to establish the relationships. However, we should confirm that data is amenable for hypothesis tests. In many situations it may not be so as the level of measurement may be lower than required or it may not follow the expected distribution.

EDA also ensures that data is having the expected coverage. In large Analytics initiative, data would be sourced from many sources. Hence, even while enough precautions are taken, there can be missing areas. For example a typical Marketing Mix analysis would have sales data from many regions or stores and numerous promotion programs. Since the sources are different, some data may miss inadvertently. When the analyst shares the EDA result with business managers, such missing will be noted.

EDA helps us to understand the possible impact of variables on the problem we are examining. Analytic process will formally establish this. But EDA will indicate possible candidates. Hence, it could be a tool for prioritizing the variables in order of importance. Since a typical analytics initiative deals with numerous variables, this is important.

EDA also will establish nature of relationship between variables and the problem. We may find that relationship between a variable and the target is not linear as expected. Such a situation calls for data transformation or change in the model estimation technique. Similarly, we may find that the relationship is not uniform as expected. This would require pre-processing the variable to categories in some cases.

In general, EDA is divided into univariate and bivariate analysis. In the univariate, our interest is to evaluate each of the variables independently. The objective is to understand the nature of the variable by examining mean, variation, skewness etc. if the variable is continuous. If the variable is categorical, we will evaluate the number of levels and the spread of each of the levels. In the case of bivariate analysis, each variable is evaluated against the target variable to get an early lead on the importance of the variables and the nature of relationship.

Analytics in Practice

Internet of Things and Analytics

Internet of Things refers to a network of everyday objects (or industrial objects) connected to internet that can share information and complete tasks without being intervened by humans. These will consists of numerous sensors and objects that will generate massive amount of data that can be analyzed to improve our quality of life in many ways.

For example, consider a refrigerator that will track temperature and humidity of all compartments and also consider the electricity rate at various time of day to activate the compressor. Similarly, a sprinkler system may consider the water requirement of the lawn through the sensors, weather forecasts and water rates to decide the watering of your lawn. It is a growing field and the expectation is very high. According to Intel, cities will spend $41 trillion in the next 20years for infrastructure for IoT[1].

Analytics is in the picture as it will help to analyze this huge data stream to take optimal decisions. We may resort to advanced approaches like stream processing, in-memory analytics, cloud analytics etc. to leverage this. Interestingly, it will also pose a challenge to Analytics practice as it will require new approach in terms of tools and techniques.

Business organizations are spending huge amount of resources to leverage this technology and bring the benefit to the people. Some of the expected benefits are:-

- Improved security of home. The sensors will keep track of the movements and let the owner know about suspicious activity or inactivity. This especially useful in elder care.

- Improving the performance of all public infrastructures like road and air traffic, Electricity, Water etc.

    - Electric grids can be made smart. Information from various sensors will help utilities to design smart meters, connect renewable resources and improve system reliability. It will be able to interact with electric vehicles to use these as a temporary storage of energy to increase overall efficiency.

    - Public water supply can be improved by collecting data using sensors about supply, demand, leakage, quality etc. (take a look at this organization that is already leveraging some of these ideas http://www.takadu.com/)

    - Transport solutions based on IoT can speed up traffic flows, reduce fuel consumption, prioritize vehicle repair schedules, and save lives. It can even price auto insurance premiums optimally.

[1] http://www.nationmultimedia.com/technology/Global-tech-giants-fully-committed-to-Internet-of--30247395.html

4.1 UNIVARIATE ANALYSIS

There is no single technique that can be applied for univariate analysis. Depending on the nature of the variable, we will choose appropriate techniques. Packages like SAS (Proc Univariate) and SPSS (DESCRIPTIVES) provide results helpful for univariate analysis through a single command. Analysis approach differs significantly based on whether the variable is continuous or categorical.

Qualitative (Categorical) Data

For categorical variables, we will be interested in the number of levels and the corresponding frequency. Table and the chart below shows a typical representation using the example of occupation of the customer of a business.

Occupation	Count	Percent
Govt. Service	5800	23%
Pvt. Service	6500	26%
Entrepreneur	2600	10%
Professional	4800	19%
Educator	5300	21%
Total	25000	100%

Table 4.1 Counts and Percentage of Occupation Categories

The table shows that there are 5 levels in occupation and the proportion of ‘Entrepreneurs’ are less than other occupations. Since each of the levels are populated fairly well, this doesn’t pose any problem. However, an analyst should be prepared for following issues.

- Some of the levels could be weakly populated. For example, in the data above, entrepreneur may be only 1% of the sample. In such a situation, it may not impact the dependent variable merely because of low coverage. This occupation could be merged with similar occupation for further analysis.

- Another issue typical of Analytics is too many levels. Many a times analyst ignore this variable as too messy to have any value. However, it could be a valuable variable if right type of pre-processing is applied. An example could be zipcodes of the address. Analyst can look for background information of the zipcodes (eg:- percapita income, literacy, distance from large town etc.) that are relevant and merge to the database and then analysis can be done on this basis. Hence, this approach is to leverage external information to add value to the data which was otherwise meaningless.

The chart below present this data visually.

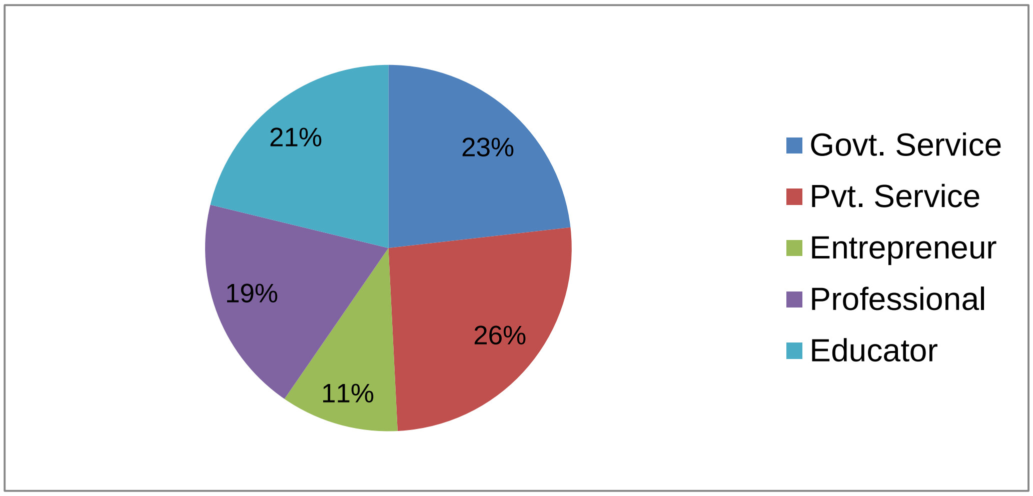


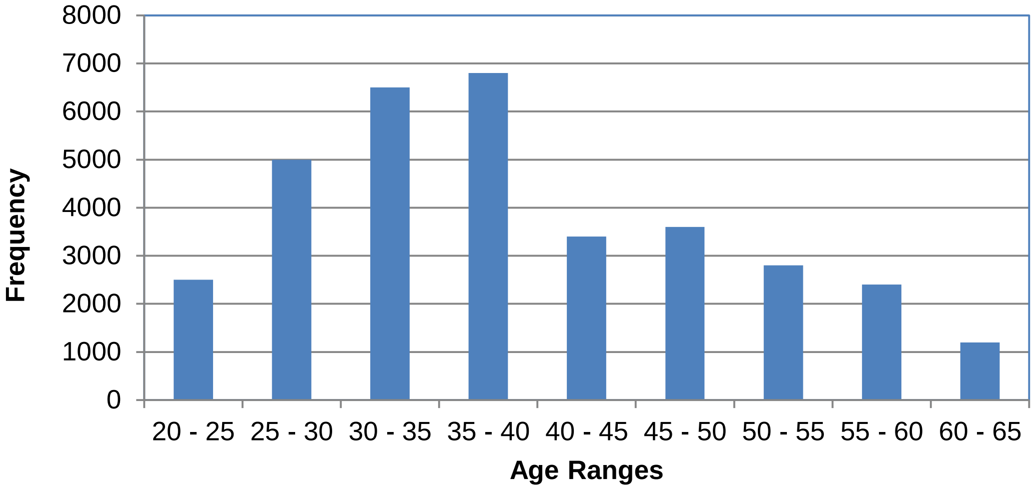
Figure 4.1 Proportion of Sample by Occupation

Quantitative Data

Many variables of interest in a typical analytics exercise will be quantitative in nature. We are interested in understanding the nature through evaluating central tendency, dispersion and distribution. Mean is commonly used as measure of central tendency and standard deviation is the commonly used measure of dispersion.

Basic approach to understanding distribution of a variable is to plot histogram of the data. Histogram is a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values of the variable. To manually construct a histogram, divide the variable into 5 to 15 groups (bins) with equal ranges which make sense (for ages it could 5 or 10 years and for income, it could be multiples of 100 or 1000). Then count the occurrences for each of the range and plot it. The number of bins should depend on the amount of data and the spread. If the number is less, it will hide the variability in the data and if it is more, the distribution could be misleading by highlighting the variation which may not have any statistical implication. In case you are undertaking any analysis which assume a distribution (like normality), it make sense to test for it and provide the result along with the chart.

The chart below shows the distribution of Age of customers of business organization. We chose an interval of 5 years to construct the histogram.



At α = 0.1, Chi-Square test indicate that the distribution of Age is not normal

Figure 4.2 Frequency by Age Ranges

The chart shows that the distribution of Age is not exactly normal. It was tested using Chi-Square and was found that the distribution is not normal.

Skewness and Kurtosis:- Two concepts closely connected with nature of distribution is skewness and kurtosis. Skewness refers to lack of symmetry of the curve. It happens when values are concentrated either on low or high end of the range.

Another behavior of interest is kurtosis. Kurtosis measure the peakness of the distribution. Larger kurtosis indicate lower variation. In such a situation the variable is similar to a constant and may not carry any information.

Normality:- Quantile-Normal (Quantile-Quantile) plots are the visual approach to check for normality. In fact the chart can be created to check the conformance of the variable with any theoretical distribution. For convenience, we will consider the case of checking for normality. By examining the plot, we can detect skewness, kurtosis and even existence of outliers.

The figure below shows the observed and expected values of a variable (Age). The expected values are the values if the distribution of the variable was exactly normal. We notice observed values are distributed on the both sides of the diagonal line indicating no serious divergence from normality.

Innovative Analytics Companies

DataKind

USING DATA IN THE SERVICE OF HUMANITY

Analytics is not all for business and profit. Organizations that don’t generate profit also can benefit immensely. But they may not be in a position to organize data or employ analysts. Datakind (datakind.org) is an initiative that tries to harness the power of data science to help such organizations.

It is a nonprofit organization devoted to connecting data scientists with organizations seeking pro-bono help. It is headquartered in New York City and has Chapters in Bangalore, Dublin, San Francisco, Singapore, the UK and Washington DC. It has already worked with organizations like Amnesty International, the Grameen Foundation, the New York City Department of Parks & Recreation, and the World Bank.

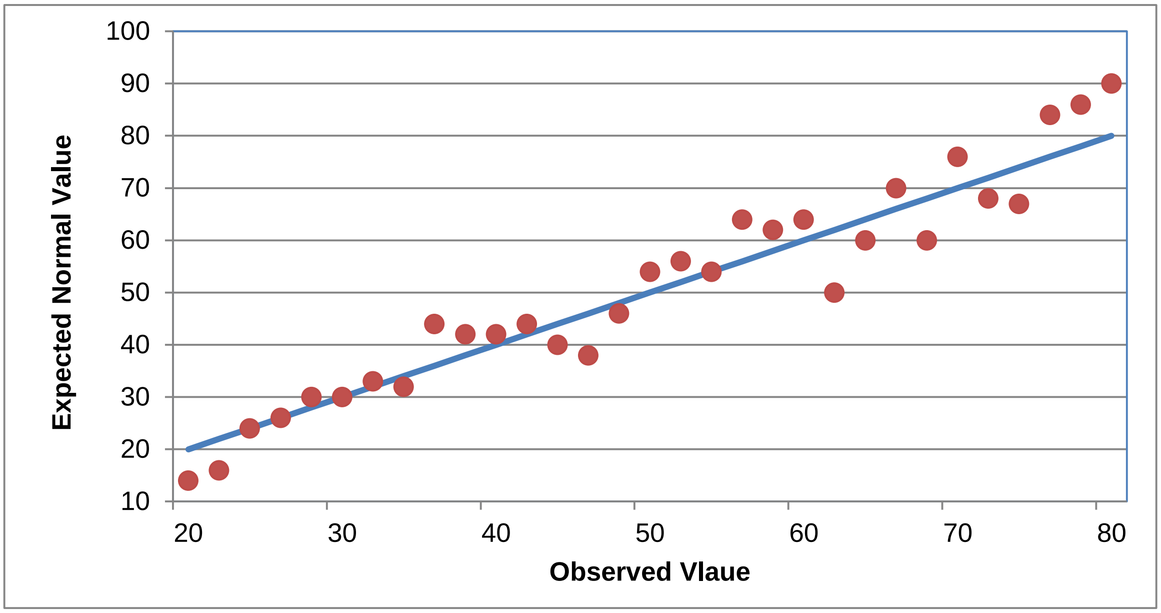


Figure 4.3 Quantile-Quantile Chart

By examining the pattern by which observed values different from expected, we can identify the reasons for divergence. Chambers et al. (1983) and Fowlkes (1987) discussed the interpretations of commonly encountered departures from linearity, and these are summarized below.

Description of Point Pattern	Possible Interpretation
all but a few points fall on a line	outliers in the data
left end of pattern is below the line; right end of pattern is above the line	long tails at both ends of the data distribution
left end of pattern is above the line; right end of pattern is below the line	short tails at both ends of the data distribution
curved pattern with slope increasing from left to right	data distribution is skewed to the right
curved pattern with slope decreasing from left to right	data distribution is skewed to the left
staircase pattern (plateaus and gaps)	data have been rounded or are discrete

Table 4.2 Interpreting Quantile-Quantile Pattern

Measures of Central Tendency:- Central tendency measures deal with middle value which is a representation of the variable. Most commonly used measures are mean and median and rarely mode. For a symmetric distribution, mean is the central point and rest of the data is distributed on both sides evenly. For non-symmetric distribution, the mean is similar to 'centre of gravity' about which the data balances.

Median is another commonly used measure. This is the middle value when the values are arranged in order. This is a powerful measure as this measure separates the population in terms of higher and lower. For example, if the median value of residence is $250000, it means that half the population resides in houses costing more than this and the other half lives in household costing lower than this. A mean of the same measure would involve houses costing tens or hundreds of millions to be included in the calculation leading to an estimate that is not representative.

An important quality of Median is the robustness of the measure. Since it depends on the relative position, it is not affected by changes in the values of the tails. Hence, in the above example, if there is a sharp drop in the values of high end homes (or low end homes), the median is still not affected. Hence, any change in high values or low values will have no effect on Median. For a symmetric distribution, median and mean is same.

Spread:- As important as central tendency of measures is the assessment of spread or variation. It is a measure of how far the values are located away from the measure of central tendency. Distributions can have same mean but very different spread. The chart below shows three distributions with same mean but different spread. Hence, to increase our understanding, along with central tendency measures, we should also measure the spread.

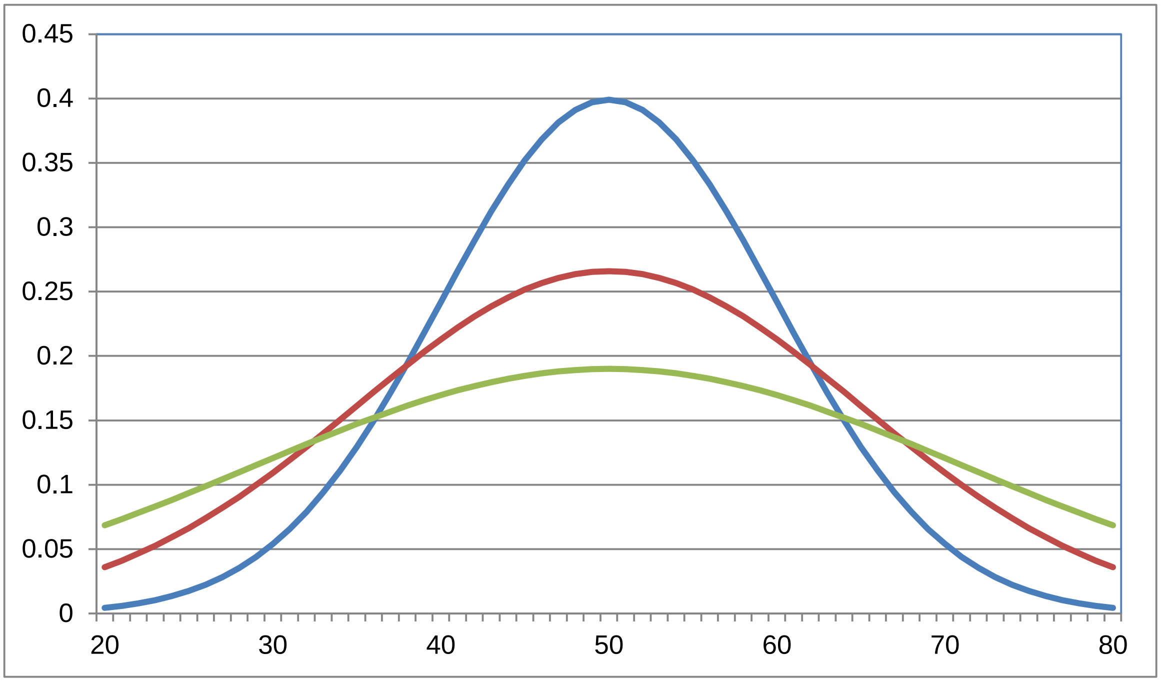


Figure 4.4 Distributions with Different Spread but Same Mean

Spread is helpful to make assessment of the reliability of the measure of central tendency. If the spread is large, the central tendency measure is not a good representation of the data. This information is useful as we will be able to take appropriate measures at analysis stage. In order to formally measure the spread, there are measures like range, inter-quartile range and population variance, the most common measure.

Population Variance:- This measure is calculated by taking a mean of the square deviation from the mean of the population. It is squared to make the deviation positive and also to give higher weightage to larger deviations. The equation below shows the case of N observations ($x_1 – x_N$) with a mean of μ.

Population variance

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

Population variance however, suffer from an important drawback. The unit of measurement is squared. For example, if the variable is in years, the variance is measured

Dependent Variable:- Univariate analysis of dependent variable must get special attention. Many a times this can provide information that will change the analysis plan. For example, consider a study of spending pattern of customers of DIY stores. Univariate analysis suggested that there are two dissimilar segment of customer class and it may not make sense to have a covering both segments.

in years squared. Hence, we take root of the variance and the term is *standard deviation*. For standard deviation the unit of measurement is same as the variable.

Sample Variance The above discussion was for population. In situations where we analyse a sample, the formula is different as given below.

Sample variance

$$s^2 = \frac{\sum(x - \bar{x}\,)^2}{n - 1}$$

Note that we divide by n-1 instead of N as in the case of population variance. Consider a case where we take many samples and has calculated the sample variance. We can prove that the mean of these variances will not be equal to population variance unless we use n-1.

As in the case of population variance, we take a square root of the sample variance. It is termed *standard deviation* and it is the most common measure of spread.

Box and Whisker Plot:- Box and Whisker plot is a common method of visually appealing presentation of the characteristics of data. It considers the quartiles, min and max values and outliers to present it on a scale. Figure below shows the components of the plot.

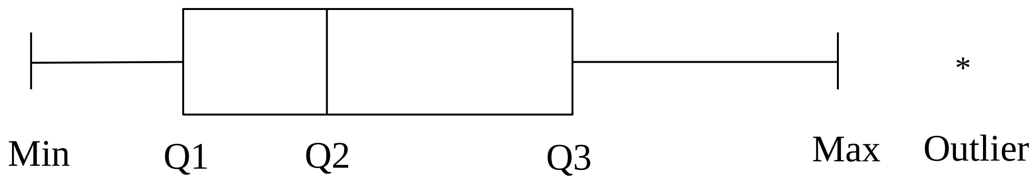


Figure 4.5 Box and Whisker Plot

Q1, Q2 and Q3 refer to the top point of the first quartile, median and top point of the third quartile. Min and Max refers of the minimum and maximum values in the database (limited to 1.5 times the interquartile range). Values outside this are considered as outliers. The advantage of this schema as evident in the figure is a detailed view of the data.

4.2 BIVARIATE ANALYSIS

The univariate analysis was helpful in evaluating the nature of variables independently. It was able to provide many valuable information that will be useful in modeling stage. The objective of bivariate analysis is to consider two variables at a time to understand how it influence each other. Most of the time, it will be a comparison between dependent variable and each of the independent variables. We will do this analysis between two of the independent variables too if the knowledge of the relationship is important.

Bivariate analysis approach too would depend on the whether these variables are quantitative or qualitative. For dependent variable, we will limit qualitative variables to cases with only two levels (binary) as this is most common situation. The table below provides the analysis type and hypothesis testing based on various types of variables.

Dependent Variable (DV)	**Independent Variable (IV)**	**Analysis Approach**	**Hypothesis Testing**
Continuous	Categorical	Means of DV by IV categories	t-test (if IV is binary) / ANOVA
Continuous	Continuous	X-Y Plot/ Correlation	ANOVA (categorize IV into groups)
Categorical	Categorical	Classification Table	Chi-square Test
Categorical	Continuous	Means of IV by DV categories	t-test (if DV is binary)/ ANOVA

Table 4.3 Bivariate Analysis Approaches

The table shows that when the dependent variable is continuous and independent variable is categorical, we can do the analysis by taking the mean of the dependent variable by the levels of the independent variable. The relationship can be tested using t-test or ANOVA. In choosing the approach, the focus should be on effective presentation of the results.

Analytics in Practice

Analytics as a Sport!!

An exciting development in the broad class of data mining and analytics is the competitions. Such competitions are open to all and provide business problem and data required to formulate a solution. This suggests that analytics too have a dimension of fun. But it is a serious stuff too when we look at the quantum of rewards.

Open competitions in the arena of data analysis or optimization has been running for many years. However, a huge interest and popularity was generated by NETFLIX when they offered a challenging problem and a shocking reward of $1M. The challenge was to improve the movie recommendation engine that Netflix uses (Cinematch). The reward was offered to any participant who can improve the accuracy of this algorithm by atleast 10% (http://www.netflixprize.com/index). Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.

The competition was started during Oct/2006 and ended Jul/2009 when a team named ‘BellKor's Pragmatic Chaos’ submitted a solution that improved the accuracy by 10.06%. In the intervening period, the journey has been quite eventful. As the teams struggled to achieve higher and higher improvement, it was concluded that there won’t be a success in the near future. Hence, Netflix decided to give yearly improvement prize too. Teams that competed decide to join and fight with higher might. At the end, there were two teams that showed more than 10% improvement, but the second team lost because they were late by 20 minutes. Another benefit of this competition was that the basic algorithm (‘collaborative filtering’) could be improved as evident in the number of articles that were published (Feuerverger, A.,2012, Bell R.M.,2010).

The tech world has noticed the excitement around this competition and there are few sites that are dedicated to such competitions (kaggle.com, crowdanalytix.com etc.). Numerous competitions are now running from companies like GE, Walmart, dunhumpy etc. One of highest prize money has been $3M on kaggle.com. These competitions offer a chance for professionals to sharpen their analytic skills and to compare it with the best. Also it offers opportunities to budding professionals to learn the basics and advance.

Robert M. Bell, Yehuda Koren and Chris Volinsky (2010), "All together now: A perspective on the NETFLIX PRIZE", Chance 23 (1): 24, doi:10.1007/s00144-010-0005-2

Andrey Feuerverger, Yu He, and Shashi Khatri (2012), "Statistical Significance of the Netflix Challenge", Statistical Science 27 (2): 202–231, doi:10.1214/11-STS368

Continuous Dependent Variable

It is quite usual to have continuous variable as dependent variable in Analytics. It could be sales, salary, loan amount, etc. etc. Such variables will be required to contrast with independent variables that could be continuous or categorical in nature. We will consider the situation of categorical independent variable first.

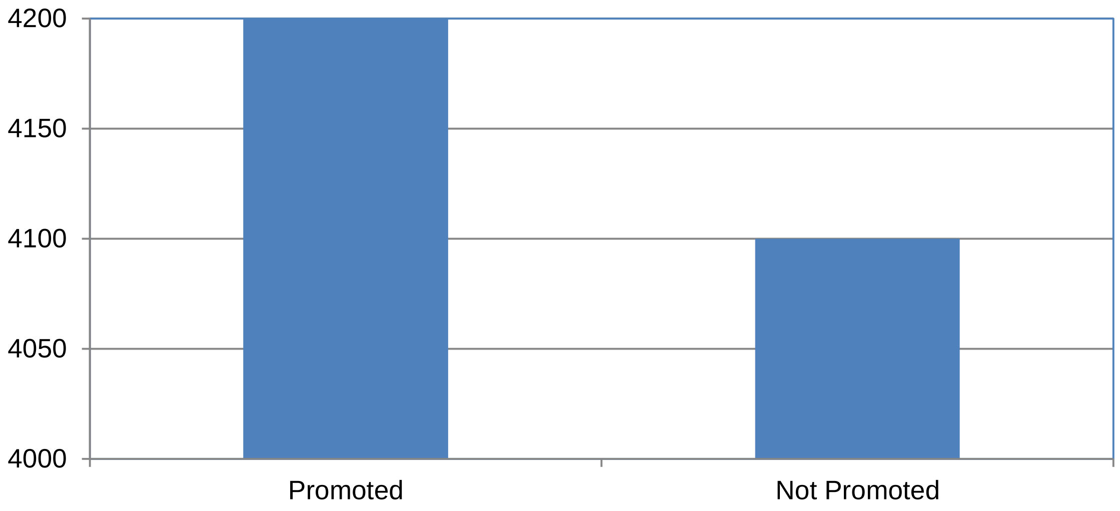
Categorical Independent Variable:- Examples of categorical independent variables include gender, home ownership, promotion, satisfaction with a product etc. It could be binary or with more than two levels. If it is binary, it makes sense to calculate mean of the dependent variable by the two levels and present as a chart. Appropriate hypothesis test for such a scenario is t-test.

Null Hypothesis:- $\beta_1 = \beta_2$

Alternate Hypothesis:- $\beta_1 \neq \beta_2$

Where β_1 & β_2 are means of dependent variable at the two levels of independent variable ('1' and '2').

The result of the t-test could be presented in the chart as shown below. The chart shows the result of an analysis of Salary as the dependent variable and promotion (binary) as the independent variable. Details of the test could be included in the appendix of your report.



A t-test was conducted and the salary differs significantly between promoted and non-promoted group ($\alpha = 0.01$)

Figure 4.6 Mean Salary of Promoted and Not Promoted Employees

If the categorical independent variable got more than two categories, it can be handled similarly. However, the appropriate hypothesis test is ANOVA.

Null Hypothesis:- $\beta_1 = \beta_2 = \beta_3 = \beta_4 \ldots$

Alternate Hypothesis:- $\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \ldots$

Where β_1 , β_2 etc. are means values of dependent variable at the different levels of independent variable (‘1’, ‘2’, ‘3’ … indicate different levels of variable).

The table below gives the output and corresponding chart of a situation where the dependent variable (salary) is continuous and one of the independent variable (department) is categorical with 4 levels.

Department	**N**	**%**	**Mean Salary**
Marketing	100	59%	4400
Finance	25	15%	4200
Operations	30	18%	4100
Human Resources	15	9%	4150
Total	**170**	**100%**	**4296**

Table 4.4 Mean salary by Departments

The chart below shows the visual presentation of this result. Note that in choosing the presentation approach, the analyst can adopt any innovative means.

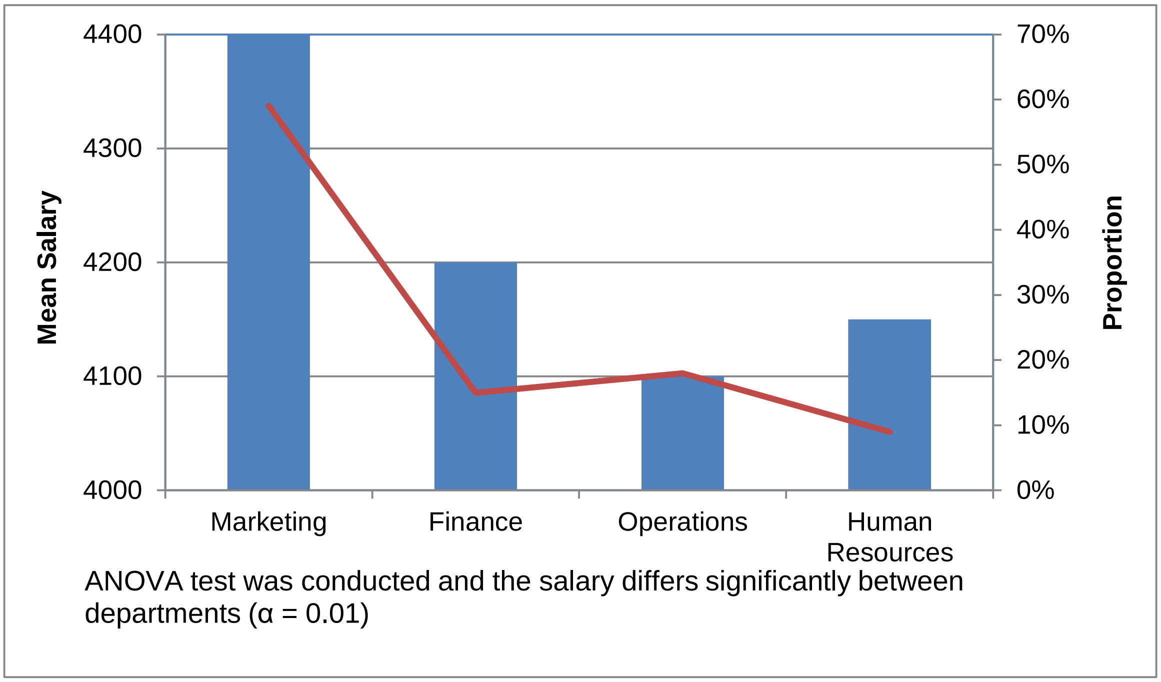


Figure 4.7 Means Salaries at Department Level

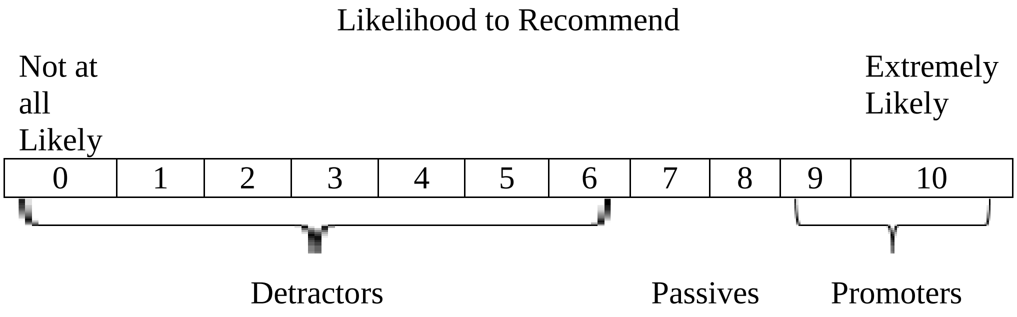
The advantage of these analyses is that these give an early view of influencers of the dependent variable. The associated hypothesis test support (or reject) the assumption we form on the basis of the chart. Hence, an important take away in the above example is that department might turn out to be an important predictor of salary in the main analysis. It also would indicate the levels that may be required to merge with others based on the coverage in the data. In the above example HR constitute only 9% of the population and Finance only 15%. Hence, HR may be combined Finance for the next level of analysis.

Tips on Technique

Net Promoter Score

Net Promoter Score is an interesting example of a simple set of data and even simpler analysis providing valuable (and even strategic) information. The score is being used to gauge the loyalty of a firm's customer. It is considered as an alternative to traditional customer satisfaction scores. Various researchers have found that it is correlated with revenue growth.
The score was developed by Fred Reichheld, Bain & Company, and Satmetrix. It was introduced by Reichheld in his 2003 Harvard Business Review article "One Number You Need to Grow". It is based on a direct question: *How likely is it that you would recommend us to a friend or colleague?* The scoring for this answer is most often based on a 0 to 10 scale. Promoters are those who respond 9 or 10 and are considered loyal enthusiasts. Detractors are those who respond of 0 to 6 - unhappy customers. Scores of 7 and 8 are passives, and they will not directly affect the formula. NPS is calculated by subtracting the percentage of customers who are Detractors from the percentage of customers who are Promoters.

How likely is it that you would recommend us to a friend or colleague?



0	1	2	3	4	5	6	7	8	9	10

Net Promoter Score = %Promoters - %Detractors

NPS can be as low as −100 (everybody is a detractor) or as high as +100 (everybody is a promoter). An NPS that is positive is considered as good, and an NPS of +50 is excellent.

Reichheld, Frederick F. (2003), “The One Number You Need to Grow,” Harvard Business Review, vol. 81, no. 12 (December), 46-54

Continuous Independent Variable :- Examples of this include income, distance, turnover, age, etc. This situation can be handled by plotting the behavior supported by a correlation analysis. It is always advisable to plot this as X-Y chart as the values will be automatically scaled. Trendline may added to indicate the linear relationship between the variables.

The figure below shows the relationship between sales and advertisement.

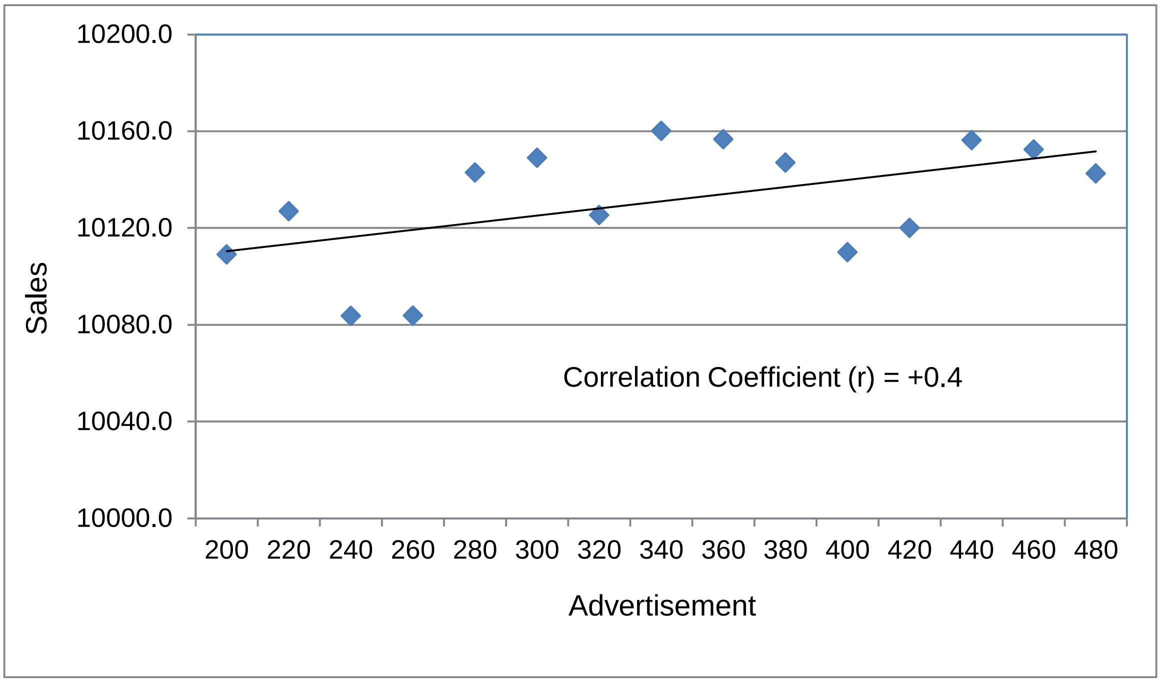


Figure 4.8 Relationship between Advertisement and Sales

It shows a weak relationship between the two as the points are distributed haphazardly around the estimated line. Correlation coefficient in the above example provides the strength of linear relationship between sales and advertisement. The value of 0.4 suggests that strength of linear relationship is weak and positive (since 1 is the highest).

If the data is time series, it is more effective to show both dependent as well as independent variable on time axis. This will show how the dependent and independent variables move together over a period of time. Hopefully this side by side comparison will provide the influence of each other. The chart below shows the mean sales and price of a product for a cluster of stores.

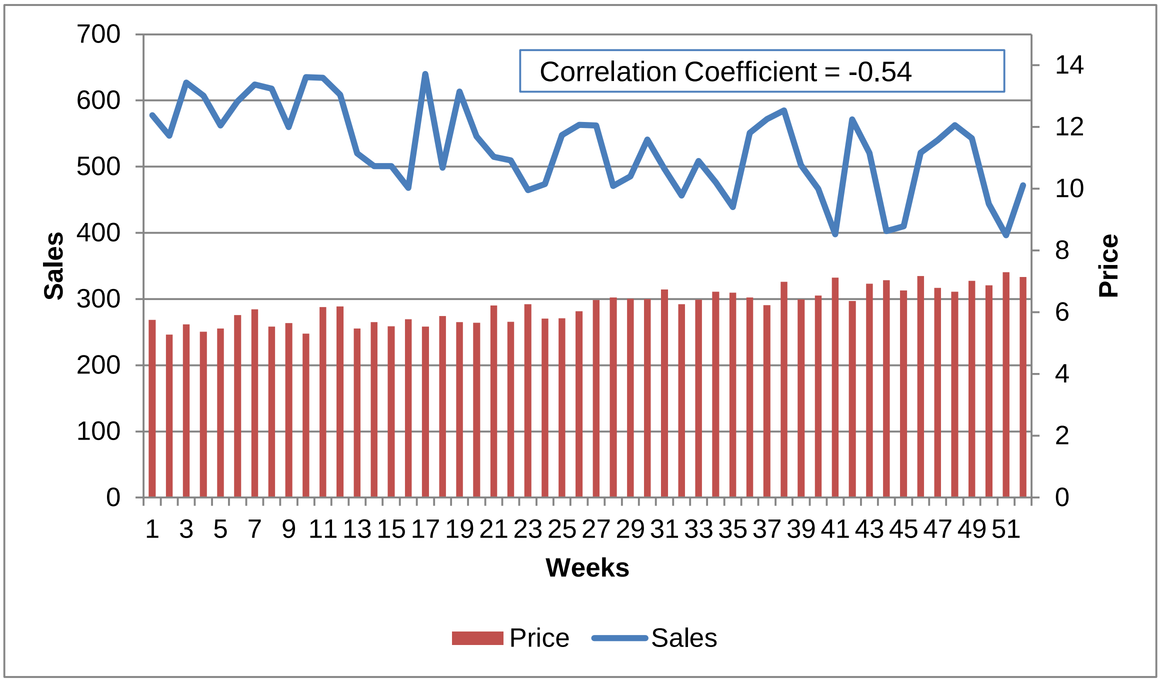


Figure 4.9 Sales and Price by Week

It shows that there is a slow trend of increasing price and decreasing sales. Between price and sales, correlation coefficient is -0.54 which indicates that 30% of the variation in sales is explained by price.

Categorical Dependent Variable

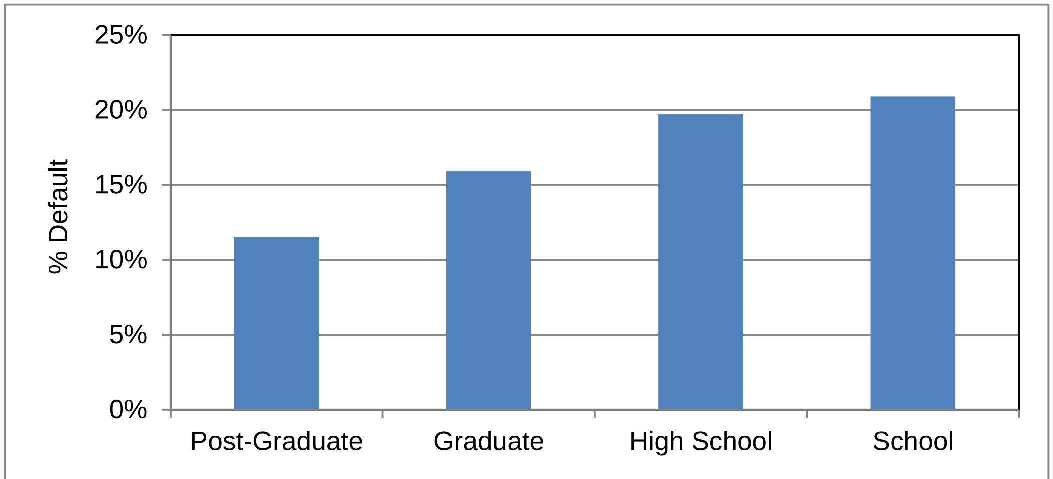
It is very common in Analytics to have categorical dependent variable. Examples include the default of a customer, purchase of a home, resignation of an employee, switching the service by customers etc. Such occurrences are quite important to business and hence, a subject of Analytics exercises.

Categorical Independent Variable:- When the independent variable is categorical, it can be presented as a table supported by a hypothesis test using chi-square. Consider a case when we are evaluating customer default with Education as the independent variable. The table is provided below.

Customer Status	Post-graduate		Graduate		High School		School		Total	
	#	%	#	%	#	%	#	%	#	%
Defaulted	15	11.5%	34	15.9 %	68	19.7 %	11 2	20.9 %	229	18.7 %
Not Defaulted	115	88.5%	180	84.1 %	278	80.3 %	42 3	79.1 %	996	81.5 %
Total	130		214		346		53 5		122 5	
Chi-square test was conducted and the independence of Default and Education was rejected (α = 0.05) (Default and Education are dependent).										

Table 4.5 Default by Level of Education

The chart below shows a possible approach to visual presentation of the above result.



Chi-square test was conducted and the independence of Default and Education was rejected (α = 0.05) (Default and Education are dependent).

Figure 4.10 Default by Education

The chart shows that default rate decreases with education. As expected, the hypothesis test confirms the dependence between education and default.

Continuous Independent Variable:- Similar to this situation is the case when we have continuous variable as one of the independent variables. The behavior can be checked by taking mean of this variable by the level of dependent variable.

Consider the same case where the customer default is being evaluated. This time, independent variable is Household Income. The chart below gives the mean Income by default. This gives an over view of the how the groups differs in terms of household income.

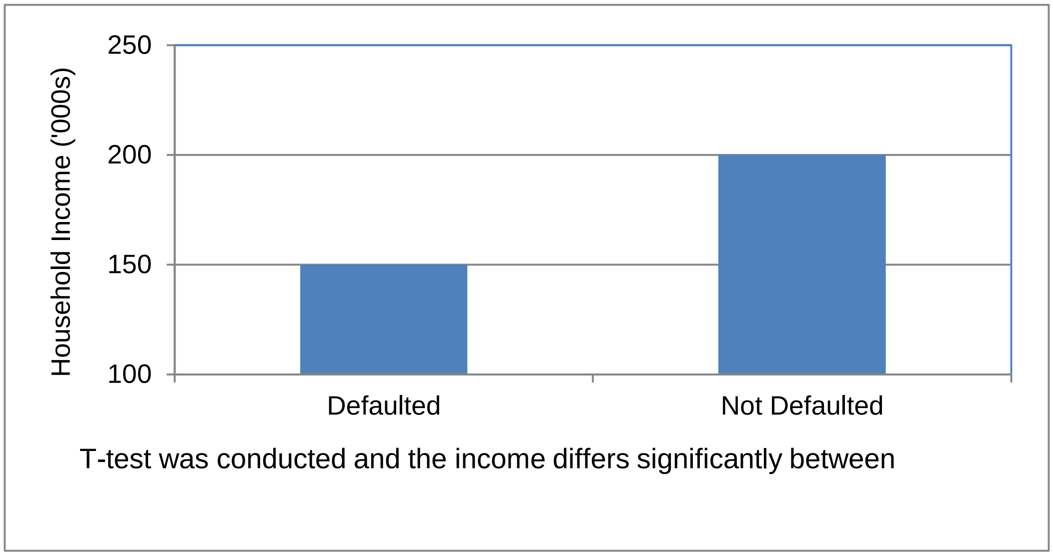
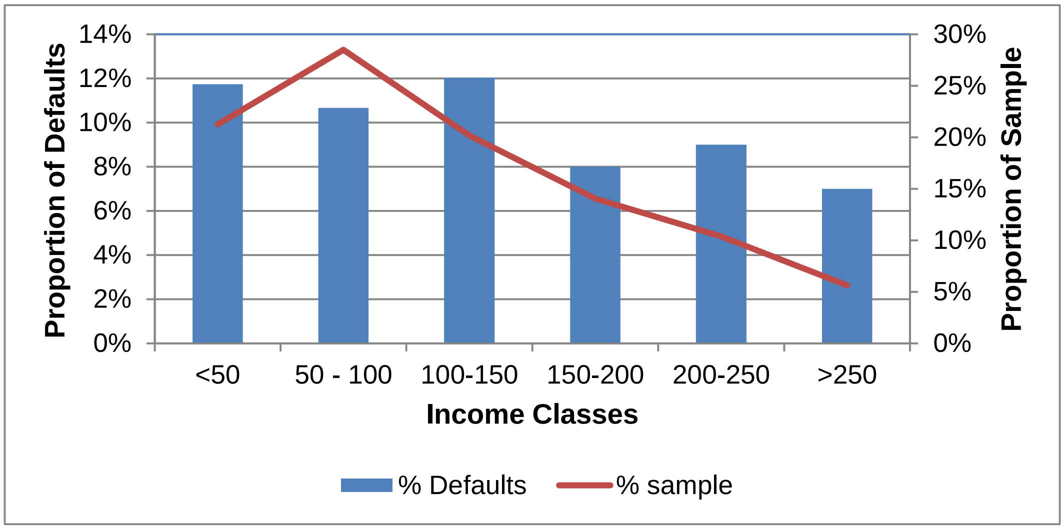


Figure 4.11 Mean Income by Default

The chart indicates that there is significant difference in household income between the two groups and it was supported by hypothesis test.



In addition to the overall difference, we are also interested in the change in default rate at various levels of income. This will help us to understand if there is break in relationship or whether it is uniform. For this exercise, we divide the income into different segments and check the default rate as below.

Figure 4.12 Default Rate by Income Classes

The chart shows that default proportion decreases as the income increases. However, we notice a break in the behavior above income level of 150 as the default rate sharply decreases above this income level. It also shows proportion in the sample. This information will help us to decide if any of the classes need to be merged. For example, '>250' class form only 9% and hence, this class may be merged with '200 – 250'. Moreover at modeling stage, we may decide to make the Income variable into binary (<=150) and (>150).

Analytics in Practice

Data Visualization – Best Practices

Data visualization; also referred as visual analytics is the application of creative visualization technology for data communication. There is a realization that even while the analysis is simple, creative visualization can bring out deeper meanings in the data. Hence, it can be considered as part of Exploratory Data Analysis. The field of visual analysis currently in a fast growth mode. There are many software packages and service providers dedicated to this. Some of these are Tableau, SAS/Visual Analytics, Qlikview, Spotfire etc.

There are few developments in the public domain that fueled interest in the visualization technology. One of that is the seminal presentation by Prof.Hans Rosling[1]. This TED presentation really broadens the horizon about what great visualization combined with passion about the subject can achieve. He also setup ***gapminder.org*** that brings together information from world over to answer critical questions. It's objective is to promote sustainable global development and achievement of the United Nations Millennium Goals by increased use and understanding of statistics and other information about social, economic and environmental development at the local, national and global levels.

Data visualization is not always an electronic presentation. There will be many situations when the mode of communication through a printout. The charts and graphs presented in *The Economist* (www.economist.com) may be considered as world standard in creative data communication.

[1] https://www.youtube.com/watch?v=hVimVzgtD6w

References

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth International Group.

Fowlkes, E. B. (1987), *A Folio of Distributions: A Collection of Theoretical Quantile-Quantile Plots*, New York: Marcel Dekker.

Tukey, John (1977). Exploratory Data Analysis (Addison-Wesley)

Questions

1. A consumer banking business is interested in building a model to predict delinquency of applicants. They are planning to use such a model to approve loan applications. It has collected application data and the record of delinquency of its customers for last few years. DSCH04TRETDLQW.xls contains this data. Conduct an exploratory analysis to understand the data and identity the possible influencers. Note that this data contains missing values and outliers and hence, require to treat it before the analysis.

2. A leading automobile business organization is troubled by attrition at managerial level. The organization is taking help of Analytics to understand drivers behind this and collected data about employees who left the organization in the last few and current employees with atleast 1 year experience. The data is contained in the file DSCH04TRETATTB.xls. Conduct and exploratory analysis and prepare a presentation for top management. The presentation should lay out the data that was already collected, initial indications on the drivers of attrition etc.

3. A leading credit card business organization sends mails to existing customers on various new products. So far the organization has been following a policy of ‘carpet bombing’ ie send mail to all customers. The management is very keen to reduce the cost of mailing and also not to contact customers too frequently. It is trying to apply Analytics to target customers for next mailings. It has collected response information and customer profile of the last mailing and is available in the file DSCH04EXPLRESW.xls. Conduct and exploratory analysis and prepare a presentation. Note that this data contains missing values and outliers and hence, require to treat it before the analysis.

Research Questions

1. Review few articles of *The Economist* and make a list of graphs in the articles. Prepare a note on the type of charts, variables involved and the reasons for the choice of this type of graph.

2. Evaluate functionalities of few visual analytics softwares available in the market.