International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

A Comprehensive Approach to Visualize Industrial Data Set to Meet Business Intelligence Requirements using Statistical Models and Big Data Analytics.

Baby.D.Dayana, Aindrila Samanta, N. Ranganathan, Kiran Venkatachalam, Neketa Jain

Abstract*: *Traditional business intelligence solutions are slower and less efficient in comparison to Big Data Analytics. Moreover, with the technology rapidly growing and reaching more number of people every day, the amount of data getting accumulated is increasing significantly. Consequently, reformation of industrial tabular datasets that are enormously huge in size into charts and graphs that provide statistical insights into the data is an important process in order to make intelligent business decisions and understand trends and patterns. This paper introduces an approach that could be utilized to carry out the above mentioned process for answering any type of business intelligence question by performing analytical techniques like Regression, Clustering, Classification and Association. For instance, tabular datasets that contain attributes of a certain object as columns would require a statistical analysis that could measure the dependability of one or more variable on the other variables and/or the relationship between the variables, in order to study the object better. We have studied the process of constructing outcomes out of raw data and deduced a series of steps starting from the collection of data to presentation of the output of models and algorithms being applied on the data that could be executed to enhance business strategies and understand data better.

***Index Terms*: Exploratory Data Analysis, Regression testing, Statistical models, Visualization.**

I. INTRODUCTION

With the escalating preponderance of data centric business models and processes, metamorphosis of raw data into useful comprehensible information is of utmost importance in today's times. Data visualization plays a critical part to representing heterogeneous, diverse and multi –dimensional raw data in an eye pleasing and easy to understand methodology, that is otherwise time consuming and tedious to read, let alone understand. This paper proposes novel sampling methods in accordance with big data visualization algorithms to avoid the redundancies occurring in visualization of data sets.

Revised Manuscript Received on 30 March 2019.
* Correspondence Author
Baby.D.Dayana, Asst. Professor, CSE DEPT ,SRM Institute of Science and Technology,Ramapuram, Chennai, Tamil Nadu, India
Aindrila Samanta, CSE DEPT ,SRM Institute of Science and Technology,Ramapuram, Chennai, Tamil Nadu, India
N.Ranganathan, CSE DEPT ,SRM Institute of Science and Technology,Ramapuram, Chennai, Tamil Nadu, India
Kiran Venkatachalam, CSE DEPT, SRM Institute of Science and Technology,Ramapuram, Chennai, Tamil Nadu, India
Neketa Jain, CSE DEPT ,SRM Institute of Science and Technology,Ramapuram, Chennai, Tamil Nadu, India

Further to understand underlying relationships between data points and in pursuit of reduction of lag in data visualization, data reduction and compression techniques have been enforced to optimize the visualization of big data set. As of here, we are handling highly diverse, heterogeneous and multi-dimensional data sets, there are consequent challenges that are usually accompanied.

To combat these challenges, Dimensionality and extension algorithm are applied to reduce the number of attributes in the data set to obtain condensed form of the same. As data volumes and variety grows, there is a significant and proportional rise in the data points in the data set, hence, data point reduction techniques have been proposed in this paper which only stores mock-up parameters instead of actual data. In today's digital era, insights procured from Exploratory Data Analysis (EDA) makes its way in strategic business and decision making. Analysts, these days, require Exploratory Data Analysis as an aid in order to train predictive models that find its application in strategic business decisions. With the rise of big data, data cleansing has become more important than ever before as far as data visualization is concerned. Every industry, be it, banking, healthcare, retail, hospitality and education is now navigating in a large ocean of data. And as the data pool is getting bigger, the probability of things going wrong too are getting higher. We further go on to explore models and algorithms like Regression testing, Correlation testing, ANOVA testing, Time-series and Descriptive statistics and their role and importance in data visualization.

II. RELATED WORK

Researchers are putting their best step forward to understand the variety of techniques that can be carried out to improve data quality. There are many aspects, but, data management and visual analytics are two main areas of research where scientists have spent their maximum time and energy. As far as data management is concerned, researchers have proposed numerous methodologies for checking, rectifying anomalies and inconsistencies in data science. Existing methodologies can be categorized into three main classifications, namely, protocol based detection techniques for data cleansing (Abedjan et al., (2015); Gschwandtner et al., and Erhart et al., (2018)), quantitative error detection techniques (Dasu and Loh, 2012) and duplication detection of data items (Elma garmid et al., 2007).

However these do not provide a unique amalgamation of data cleansing techniques, numerosity reduction techniques and data point reduction techniques which can be applied to SQL – based database. Data cleansing techniques according to Van den Broeck el al., 2005, are screening stage diagnose stage and correction stage. Data visualisation process can be further enhanced using the data cleansing techniques in accordance with the two proposed algorithms, i.e., numerosity reduction technique and data reduction technique. Exploratory data analysis is another data exploration and visual analytics aid that has revolutionized how data is studied and perceived by data analysts.

A comprehensive survey of recent advancements in the emerging field of exploratory data analysis is presented and investigated (Aindrila Gosh et al., 2018). DEEPEYE, a data visualisation system that automatically generates the corresponding visuals from a given set of data (Yuyu Luo et al., 2018) is a step forward for efficient visualisation using neural networks. Use of neural networks in data visualisation is an effective approach to inculcate the visualisation software to mimic human intelligence.

III. METHODOLOGY

In this section, we aim to present the techniques that we have envisioned for the process of data visualization. We expect to solve the existing problem under study with Big Data analytics. Let us refer to the entire process of collecting tabular data sets to producing a data representation report as a system. The system comprises of processes that take place consecutively and produces an end result that represents the industrial tabular dataset in a better way and ultimately leads to better business decision making for the user. The following processes are elaborately explained in the following sections. The system is broken down into modules which are- Data importation, exploratory data analysis, models and algorithm and data representation. Each of the module comprises of one process and the mechanism of each module depends heavily on the output of its preceding module.

A. Data collection

In our system, we are dealing with quantitative data which means our main aim is to convert raw numbers into meaningful figures. The upcoming modules include calculation of frequencies of variables and differences between variables. Our approach to build a visualized form of tabular dataset include steps where evidence is required to either support or reject a hypothesis. Consecutively, data collection is an essential procedure since all other procedures directly or indirectly depend on the output of this process.

B. Data importation

In the technology world, the word “import” refers to the process of bringing a part of information or data from various environment into the one we are currently working. If one is using a range of systems and tools to run business, there is a need to consolidate and transfer data to analyze the data at one place. For example, you can turn a CRM data, e-Commerce data into one single set to be analyzed and any process that needs to be carried out can be done with relative ease. Data importation is a very important step, the ability to import data is very crucial in software applications as it allows one application to complement another application. Data importation in earlier days was a semi-automated process from various data bases to software applications and also between two or more software applications. The industrial tabular data sets are entered into the system for the further processing through this process. This process is basically the input of raw data. The software or the tool that is being used for analyzing the data has to have an input to work upon, as a result the flat file or the excel sheet or the SQL data that needs to be worked on, is imported into the software using commands that are supported by the software. For instance, we have studied the process using the software R and the R commands to import big data files are as follows.

1) To enter a csv file as input- read.csv() or read.csv2()
2) To enter a table as input- read.table()
3) To enter a tab separated value file- delim()

Similarly, it is very simple to work with SQL data in the R software using simple queries.

C. Data cleansing

With professional practicing guidelines being acknowledged and implemented by more and more data scientists, vital shifts in research practices are being expected. One such expected development is growing emphasis on standardization, documentation and data quality. The journey of quality assurance generally includes error avoidance, data monitoring, data cleaning and documentation. All these mentioned data quality assurance processes was done manually almost a decade back. As the amount of data, i.e., volume of data grew exponentially over the years, manual data cleansing became a tedious and unimaginable task to be carried out by a human, also, before data was cleansed, administratively inaccurate and inconsistent data would lead to false conclusions and vain investments on a colossal scale. This became a major issue as companies could not afford to waste their financial resources on areas which have been falsely concluded. Such reasons provoked data scientists to come up with automatic data cleansing techniques and algorithms that can cleans the data set before any other task is carried out using the data set. Data scientists, initially, focused on trivial problems like data repetition and data mismatch, but data cleansing nowadays is much more than that. Data cleansing in today’s times is performed interactively with data wrangling tools or as batch processing through scripting. Data cleansing techniques at their stage of inception were nowhere close to mimicking human intelligence, hence lots of important data used to get affected due to severe insufficiencies at their earlier stages. As the data scientists kept working on data cleansing techniques and algorithms, processes like data auditing, workflow specification, workflow execution, post processing and controlling, parsing, data transformation, data elimination using statistical methods are possible and improve data quality drastically before data sets can be used to visualize or draw any kind of conclusions for business intelligence solutions to formulate.Data collection has become an unavoidable task of in the industries and organizations – not only for keeping record, but to support several data analysis tasks that are critical for the decision-making process for the organizational benefits.

Even though data collection and analysis of the data seem like the most important parts of the system, maintaining data quality is the compulsory step and the most difficult one in almost every large organization. The presence of incorrect or inconsistent data can put a significant change in the results of analyses, mostly negatively affecting the output of the system. As a result, there has been various researches over the last decades on various aspects of data cleaning like, to what extent it can be automatized and the different techniques to effectively carry out the process.

In this section, we survey data cleaning methods, the types of errors one must tackle usually and the importance of data cleansing.

The different type of errors that might potentially affect the further procedures are errors while doing the data entry process, measurement errors, inconsistent data sets where the data is preprocessed and summarized before being entered into the database due to variety of reasons like reduction of complexity, integration errors from data sets that are formed by merging data from various sources.

We carried out the process under study using R and concluded that to treat a raw dataset that is just imported into the software, the basic cleansing process of that data would be-

1) Check the class of the data frame using the code- class (data).
2) Check the dimensions of the data frame using the code- dim (data).
3) Carry out the summary statistics for all the columns of the data frame using the code- summary (data).
4) Carry out visual exploratory analysis. The two methods to perform this step are histogram and box plots. Histograms are used to determine the presence of outliers so they can be removed for effective analysis in further processes. It also determines if the distribution of the data is normal, bi-modal or unimodel. Boxplots are essential because it represents the median along the first, second and third quartiles. Furthermore, it is the best way of spotting outliers.
5) Correct the errors that are identified using R commands.

D. Models and Algorithms

This is the step where the manipulation, application of formulae and modelling of the data takes place, in order to provide results that expresses the analysis carried out aiming at answering certain business decision questions. The key point here is, the testing methodology or the choice of algorithm depends heavily on the reason for the analysis and the data used. As a result, data sampling and hypothesis testing are important steps of this process. To understand what data sampling and hypothesis testing is, one needs to know what exactly hypothesis means. A hypothesis is nothing but an assertion or a statement about the state of nature and the true value of an unknown population parameter. In other words, an assumption. So, data sampling refers to a statistical hypothesis technique used to select, manipulate, and analyze a subset of data points to discover hidden patterns and trends in the larger data set. Now as we know a hypothesis is an assumption, which means it is either right or wrong. The process to test this hypothesis is known as hypothesis testing. There are several types of hypothesis tests like simple hypothesis test, complex hypothesis test, null hypothesis test, alternative hypothesis test, statistical hypothesis test, parametric hypothesis test and non-parametric hypothesis test.

Let us understand the concept of hypothesis with a simple example. Consider a scenario where a marketing manager must decide whether to launch a new product or not. On analysis, the manager could arrive at the following decision: The product will be launched if the company gets a market share of 15% or more. Prediction of such outcomes that would maximize profit at minimum risk depends on hypothesis testing.

In a simple hypothesis, there exists a relationship between two variables; one is called an independent variable or cause and the other is called a dependent variable or effect.

Example: Given total Population = 100, Total No. of Male = 50, Total No. of Female = 50, Ho: μ=50

A complex hypothesis refers to the prediction of relationship between two or more independent variables or two or more dependent variables.

Total Population (μ1) = 100, No. of Male (μ2) = 50, No. of Female(μ) is HO: μ = μ1- μ2 =50

A Null Hypothesis is usually a hypothesis of "no difference." It is denoted as H0.

Null Hypothesis is performed for a possible rejection under a true assumption and always refers to a specified value of the population parameter, such as μ.

Example:

The population mean is 100 Or H0: μ = 100

An alternate hypothesis is complementary to the null hypothesis. It is denoted by H1. Alternate hypothesis is used to decide whether to employ a one-tailed test or two-tailed test.

Example:

For H0: μ = 100, the alternative hypothesis could be: H1: μ ≠ 100 or, H1: μ > 100 or, H1: μ < 100

A statistical hypothesis is a method of statistical inference performed using data from a scientific study.

Example:

Given, total no of cities = 10 Mean population (μ) = 75 H0: μ = 75

A parametric statistical test is one that makes presuppositions regarding the parameters (defining properties) of the population distribution(s) from which a individual's data is drawn. In this paper, the types of parametric tests being studied are Z-Test and T-Test; and ANOVA Test. The Z-Test and T-Test are used when two population means or proportions are compared and tested. The ANOVA test is used when equality of several population means is tested. Z-Test is performed in cases where the test statistic is t, σ is known, the population is normal, and the sample size is at least 30.

The formula to calculate z (standard statistic) is: (X- μ)/ (σ/ square root of n)

Where, n: Sample number, X: Sample mean from a sample X1, X2, …, Xn, μ: Population mean, σ: Standard Deviation

T-Test is performed in cases where the test statistic is t, σ is

unknown, sample standard deviation is known, and the population is normal.

The formula to calculate t is:

$t = (X - \mu)/ (s/\text{square root of } n)$

Where, n: Sample number, X: Sample mean from a sample X1, X2, …, Xn, μ: Population mean σ: Standard DeviationAlong with hypothesis testing and data sampling, the other data analytics techniques that are important to carry out on the data in order to make business enhancement decisions, are regression analysis, classification, clustering and association technique.Regression analysis is used to estimate the relationship between variables. Simple regression considers one quantitative and independent variable X to predict the other quantitative, but dependent, variable Y. Multiple regression considers more than one quantitative and qualitative variable (X1…XN) to predict a quantitative and dependent variable Y. R squared and adjusted R squared are important measures of a regression model and explain the variance with the help of independent variables. Factor analysis and principal component analysis are the methods used to decrease the number of variables or factors in a model. Multiple regression has two types of models; linear and non-linear.Classification is a technique to determine the extent to which a data sample will or will not be a part of a category or type. The classification process uses two techniques for prediction: model construction and model usage. Different classification techniques include logistic regression, support vector machines, K-nearest neighbors, Naive Bayes classifier, decision tree, and random forest classification. Bias and Variance are the two types of major errors in a predictive model. Validation methods such as K-fold cross validation can be used to decrease over fitting in a model.Cluster analysis or clustering is the most commonly used technique of unsupervised learning to find data clusters such that each cluster has most closely matched data. K-means is a Prototype-based method for clustering that involves assigning training data to matching cluster based on similarity and using an Iterative process to get data points in the best clusters possible. Hierarchical Clustering clusters n units/objects, each with p features, into smaller groups and creates a hierarchy of clusters as a dendrogram. DBSCAN (Density-Based Spatial Clustering and Application with Noise) is used to identify clusters of any shape in a dataset containing noise and outliers. Prototype-based clustering assumes that most of the data is located near prototypes (element of data space representing a group of elements).

Association rule mining finds interesting patterns in a dataset. The interesting relationships can have two parameters: frequent item sets and association rules. An association rule is a pattern that states when X occurs, Y occurs with a certain probability. The measures of the strength of association rules are support and confidence. Apriori is an algorithm for frequent item set mining and association rule learning over transactional database. The Apriori algorithm includes two steps: mining all frequent item sets and generating rules from frequent item sets.

E. Data Representation

This is the step where the outcomes of the data analysis process are visually communicated in different ways to represent the results of the models and the algorithms that were being carried on the data. To study this process, we researched the ways of data visualization using the software R. Data visualization in R can be done using the following graphics: Bar chart, Pie chart, Histogram, Kernel density plot, Line chart, Box plot, Heat map, Word cloud.

Bar plots are horizontal or vertical bars used to show comparisons between categorical values. They represent length, frequency, or proportion of categorical values. The syntax is- barplot(x).

[1]A pie chart is a graph in which a circle is divided into sectors, each representing a proportion of the whole. The syntax is- pie(attributes).

[2]A histogram represents the distribution of a continuous variable and the frequency of values bucketed into ranges. The syntax is- hist(x).

[3]A Kernel density plot shows the distribution of a continuous variable. The syntax is plot(density(x)).

[4]A Line chart is used to represent a series of data points connected by a straight line. It helps visualize data that changes over time. The syntax is- lines (x, y, type).

[5]Box plot, also called whisker diagram, displays the distribution of data based on the five-number summary: Minimum, First quartile, Median, Third quartile, Maximum. The syntax is- boxplot(data). Word cloud (also called tag clouds) highlights the most commonly cited words in a text using a quick visualization. The syntax is- wordcloud (words= data, freq= freq, min.freq = 2,). Heat map is a two-dimensional representation of data in which the values are represented by colors. The two types of heat maps are: Simple Heat Map: Provides an immediate visual summary of information and Elaborate Heat Map: Helps in understanding complex data sets. The syntax is- heatmap (data, Rowv=NA, Colv=NA) Futhermore, ggplot2 is a data visualization package of R that provides a general scheme for data visualization. It breaks up graphs into semantic components such as scales and layers. It is an alternative for the basic graphics of R.

IV. RESULTS AND DISCUSSION

a) THE MEANS PROCEDURE

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
Order_ID	Order_ID	30	110015.50	8.8034084	110001.00	110030.00
Sales	Sales	30	152.9666667	63.1759803	33.0000000	250.0000000
Quantity	Quantity	30	3.1666667	1.2340942	1.0000000	5.0000000
Discount	Discount	30	0.0256667	0.0154659	0.0100000	0.0500000
Profit	Profit	30	72.1063333	44.6008984	3.2500000	135.6000000
Shipping_Cost	Shipping Cost	30	7.2106333	4.4600898	0.3250000	13.5600000

Fig. 1

Given a tabular data set containing retail attributes, we have executed the proposed approach given in this paper on the data set and performed descriptive analysis on the data to understand it better.

Descriptive statistics basically refers to analysis that would describe, show or summarize data in a significant way that could further help us, for instance, to identify a pattern. Although it is generally performed on raw data to move forward with other analytical tactics, it does not help us to conclude a hypothesis or to predict a trend.

b) THE TTEST PROCEDURE

Variable: Profit (Profit)

N	Mean	Std Dev	Std Err	Minimum	Maximum
30	72.1063	44.6009	8.1430	3.2500	135.6

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
72.1063	55.4521	88.7606	44.6009	35.5205	59.9577

DF	t Value	Pr > |t|
29	8.86	<.0001

Fig. 2(a)

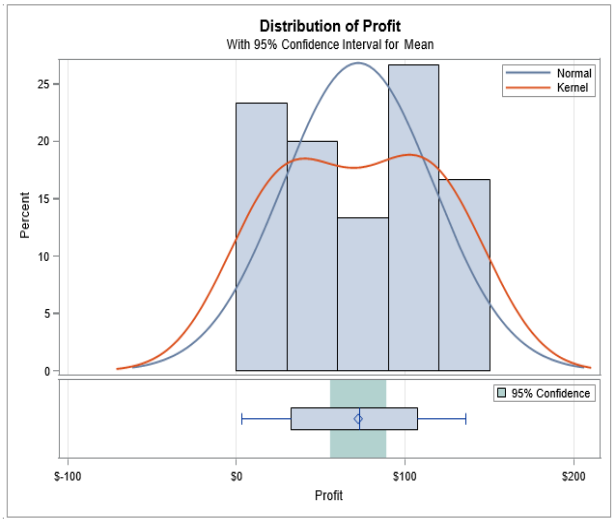


Fig. 2(b)

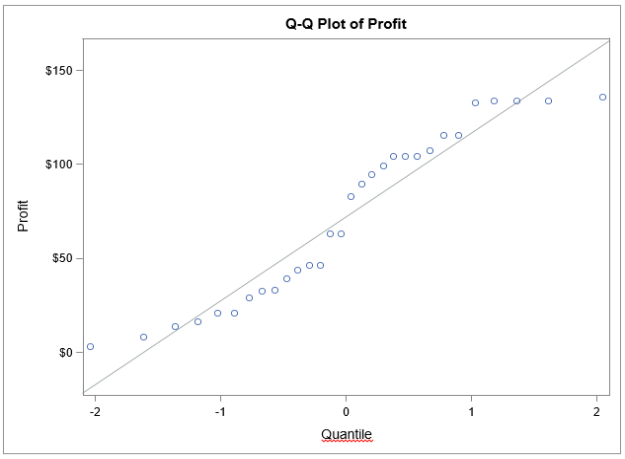


Fig. 2(c)

After performing descriptive analysis on the data set given, we move on to perform a one way T-Test procedure on the the variable profit from the data set. This procedure is used to compare a sample with a given value and the default value set is 0. Summary statistics are represented at the top of the output. The sample size (N), mean, standard deviation, and standard error are represented with the minimum and maximum values of the profit variable. The 95% confidence limits for the mean and standard deviation are displayed next. The degrees of freedom, t statistic value, and p-value for the t test are displayed at the bottom of the output. At the 5% level, this test indicates that the mean length of the court cases is significantly greater than 0. Figure 2(b) and Figure 2(c) shows a histogram with overlaid normal and kernel densities, a box plot and the 95% confidence interval for the mean. The confidence interval includes the null value, consistent with the acceptance of the null hypothesis at ALPHA=0.5.

c) THE REG PROCEDURE

Model: MODEL1
Dependent Variable : Sales

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-317.49838	111.47155	-2.85	0.0111
Profit	Profit	1	-0.70605	0.65853	-1.07	0.2986
Quantity	Quantity	1	76.21860	88.74949	0.86	0.4024
Unit_Price		1	0.59358	1.82331	0.33	0.7487
PROFIT_LOG		1	7.61138	10.37035	0.73	0.4730
QUANTITY_LOG		1	-14.02591	113.90434	-0.12	0.9034
Unit_Price_LOG		1	65.57799	29.79296	2.20	0.0418
PROFIT_EXP		1	-3.7893E-57	1.40245E-57	-2.70	0.0151
QUANTITY_EXP		1	0.88475	1.10190	0.80	0.4331
Unit_Price_EXP		1	4.61542E-46	1.73468E-46	2.66	0.0165
PROFIT_SQ		1	0.00857	0.00353	2.43	0.0264
QUANTITY_SQ		1	-2.26407	2.68423	-0.84	0.4107
Unit_Price_SQ		1	-0.00374	0.00987	-0.38	0.7094

Fig. 3(a)

Regression Test
Dependent Variable : Sales

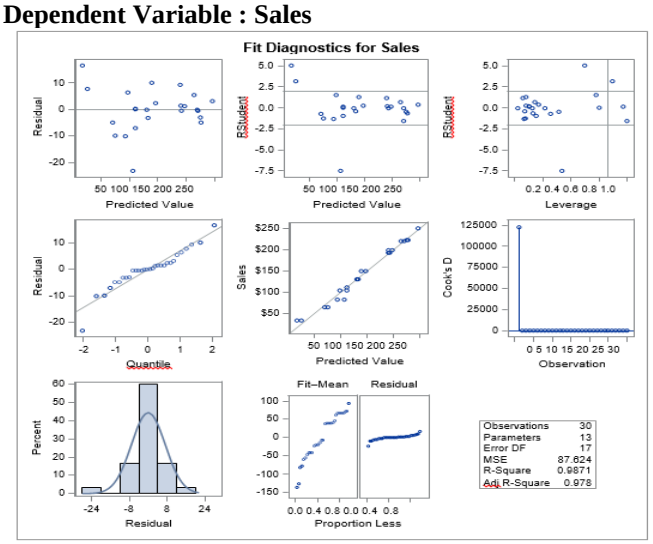


Fig. 3(b)

Regression Test
Dependent Variable : Sales

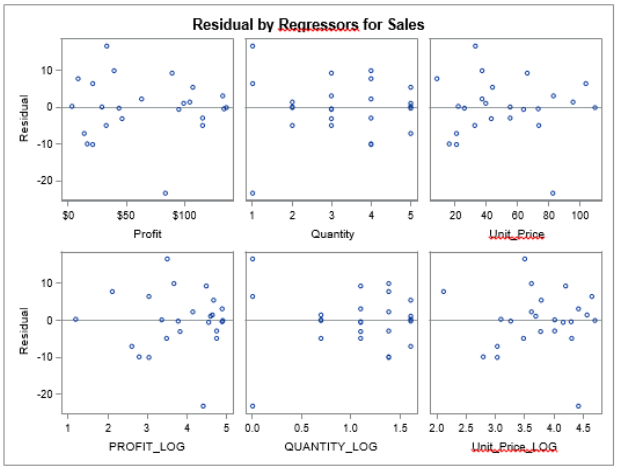


Fig. 3(c)

Figure 3 represents the "parameter estimates" of the given date set. Here, we have carried out the procedure of Regression analysis to understand the dependability of one variable Sales on other variables Profit, Quantity and Price. We have manipulated our data set in the exploratory data analysis process and extended it to add columns containing log, squared and cubed values of the existing data. Additionally, we have added a column containing price of each unit for further mathematical implications. A pattern in the residuals would suggest there is variance as far as data is concerned, which is non-constant. Figure 3(a) is helpful to indicate a slight trend in the residuals; they appear to increase slightly as the predicted values increase. The residuals arranged in a fan-shaped manner might indicate the requirement of a variance-stabilizing transformation. A curved trend (such as a semicircle) might indicate the requirement of a quadratic term in the model. Since these residuals represented in the figure 3(c) have no apparent trend, the analysis is considered to be acceptable.

d) ANOVA Test using R

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
2 observations deleted due to missingness
> summary(Aov2)
              Df    Sum Sq   Mean Sq F value   Pr(>F)
Season         7 1.778e+09 254071006    3.73 0.000589 ***
Residuals    541 3.685e+10  68109072
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
1 observation deleted due to missingness
> summary(Aov3)
              Df    Sum Sq  Mean Sq F value Pr(>F)
Material      23 1.54e+09 66977402    0.948  0.533
Residuals    525 3.71e+10 70667348
1 observation deleted due to missingness
>
>
> Model2 <- lm(TotalSales~Style + Season + Material, data = p)
> Model3 <- lm(TotalSales ~ Style + Price, data = p)
> summary(Model2)

Call:
lm(formula = TotalSales ~ Style + Season + Material, data = p)

Residuals:
   Min     1Q Median     3Q    Max
-13520  -3100  -1418    615 102954

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2973.75    5213.16   0.570  0.56864
StyleBrief        3890.31    2511.31   1.549  0.12198
StyleCasual       1199.18    1750.12   0.685  0.49353
Stylecute         2794.02    2078.33   1.344  0.17944
Stylefashion     -2039.83    8412.64  -0.242  0.80851
StyleFlare       -1305.60    6187.78  -0.211  0.83298
StyleNovelty       -50.19    2984.45  -0.017  0.98659
StyleOL          -1967.70    8448.11  -0.233  0.81592
Styleparty        -550.10    2053.13  -0.268  0.78886
```

Fig. 4(a)

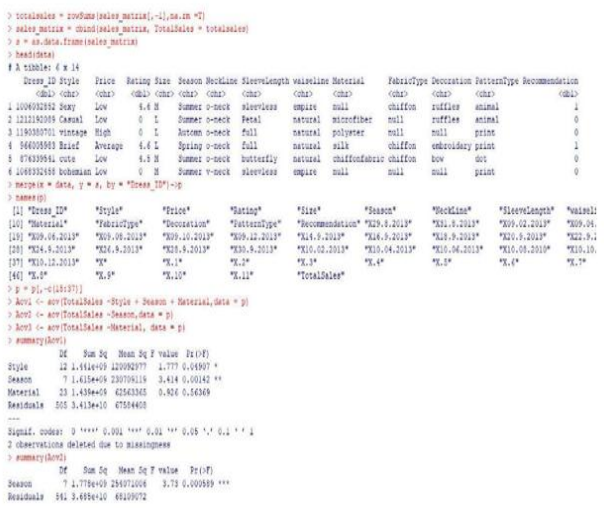
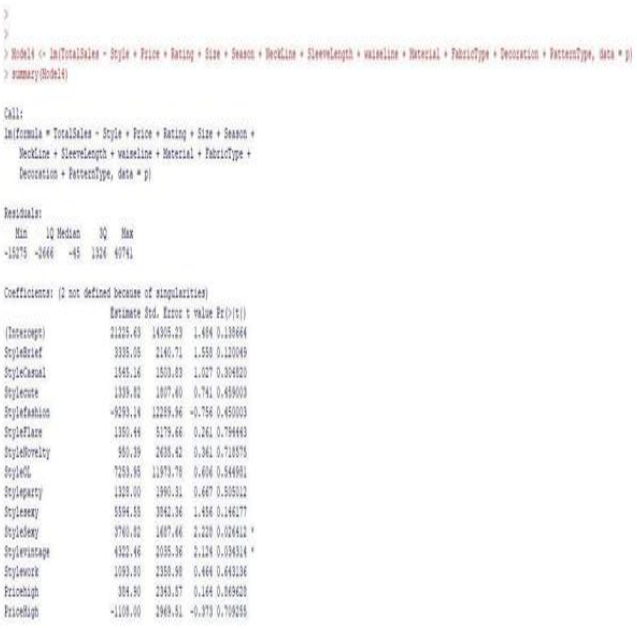


Fig. 4(b)

As we know our main aim is to answer business intelligence questions, here we have tried to find out how total sales is affected by the style, season and material of the clothes sold in the retail dress store. For this, we have executed our proposed approach step by step and used the ANOVA Testing which is a data analytic technology in the "models and algorithms" module. In Figure 4(b), we can observe that the F value of season is 3.414 and p-value is less than 0.05. Thus, we can conclude that there is a significant relationship between the season and the total sales of the retail store. The style of the dresses affect weakly and the material has negligent relationship with the total sales relatively.

e) Multiple Regression analysis using R



```
Decorationflowers          -2489.74    5012.47  -0.497 0.619670
Decorationhollowout         -936.15    2385.15  -0.392 0.694905
Decorationlace                96.21    1931.64   0.050 0.960302
Decorationnone              -163.70    4765.62  -0.034 0.972616
Decorationnull              -765.73    1832.41  -0.418 0.676258
Decorationpearls            -793.94    7416.89  -0.107 0.914807
Decorationplain            -3030.36    6731.17  -0.450 0.652813
Decorationpockets          -1160.65    3578.86  -0.324 0.745877
Decorationrivet            -1814.71    4449.17  -0.408 0.683585
Decorationruched           -3173.06    4774.26  -0.665 0.506681
Decorationruffles           2934.75    2490.58   1.178 0.239367
Decorationsashes            -196.73    2028.55  -0.097 0.922792
Decorationsequined          -622.56    2605.21  -0.239 0.811254
Decorationtassel           -2309.66    6990.98  -0.330 0.741289
DecorationTiered           12077.57    7825.36   1.543 0.123534
PatternTypecharacter       -3543.57    6661.54  -0.532 0.595062
PatternTypedot             -4152.07    2596.81  -1.599 0.110636
PatternTypefloral           -717.95    4941.37  -0.145 0.884553
PatternTypegeometric        2437.93    3757.00   0.649 0.516777
PatternTypeleapord         -8984.03    6981.96  -1.287 0.198932
PatternTypeleopard         -1576.64    3752.34  -0.420 0.674585
PatternTypenone            -3417.15    6829.68  -0.500 0.617115
PatternTypenull            -4161.62    1779.79  -2.338 0.019869 *
PatternTypepatchwork       -4504.31    1917.42  -2.349 0.019306 *
PatternTypeplaid           -5861.58    4246.33  -1.380 0.168244
PatternTypeprint           -2768.83    1803.46  -1.535 0.125511
PatternTypesolid           -2754.12    1728.95  -1.593 0.111968
PatternTypesplice          -4492.11    5425.68  -0.828 0.408205
PatternTypestriped         -3926.74    2363.66  -1.661 0.097443 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6372 on 397 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.5826,    Adjusted R-squared:  0.4259
F-statistic: 3.719 on 149 and 397 DF,  p-value: < 2.2e-16
```

Fig. 5

Let us consider a test case where to enhance the sales, the the attributes of dresses have to be analyzed to ultimately find out which are the leading factors affecting the sale of a dress. To answer this problem statement, we chose the multiple regression analysis technique in the "models and algorithms". We can interpret from the figure 5 very easily that the factors affecting the sales are style sexy, style vintage, pattern type null, pattern type patchwork and pattern type striped.

f) Correlation Test using R

To test the association between two variables total sales and rating, we opted for the Pearson's product-moment correlation technique. In the result above, the t-value us 4.1878, there are 548 degrees of freedom and the p-value is 3.281e-05 which is less than ALPHA=0.05. Thus, we can conclude that Total sales and rating are significantly correlated with a correlation coeffiency 0.1761003.

```
>
>
> attach(p)
> cor.test(TotalSales, Rating)

        Pearson's product-moment correlation

data:  TotalSales and Rating
t = 4.1878, df = 548, p-value = 3.281e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.09387604 0.25593845
sample estimates:
      cor
0.1761003
```

Fig. 6

REFERENCES

1. Shaun Bangay, "Visview: A system for the visualization of
2. Multi-dimensional data", in "Visual Data Exploration and Analysis V".
3. (TA 1505 Pse 3298)
4. Matthew Ward, "Overview of Data Visualization", from www.cs.wpi.edu
5. [7] Steven Richard Hollasch, "Four-Space Visualization of 4D Objects", from http://www.research.microsoft.com/~hollasch/thesis/default.htm
6. O. Kumar and A. Goyal, "Visualization: a novel approach for big data
7. analytics," Proceedings of the Second International Conference on Computational Intelligence & Communication Technology, 2016, pp. 121-124.
8. "Data visualization techniques," SAS,
9. http://www.sas.com/en_us/offers/sem/data-visualization-techniques-2332568.html?keyword=data+visualization+techniques&matchtype=p&publisher=google&gclid=COycgoCbutACFcolgQodqwgIiA
10. R. R. Laher, "Thoth: software for data visualization and statistics,"
11. Astronomy and Computing, vol.
12. 17, 2016, pp. 177-185.
13. X. Li et al., "Advanced aggregate computation for large data
14. visualization", Proceedings of IEEE Symposium on Large Data Analysis and Visualization, 2015, pp. 137,138.
15. S. A. Murhy, "Data visualization and rapid analytics: applying tableau
16. desktop to support library decision-making," Journal of Web Librarianship, vol. 7, no. 4, 2013, pp. 465-476.
17. [1] https://itmodes.wordpress.com/data-science/
18. [2] https://www.tutorialspoint.com/r/r_histograms.htm
19. [3] https://itmodes.wordpress.com/data-science/
20. [4] https://itmodes.wordpress.com/data-science/
21. [5] https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review