EasyChair Preprint

№ 6919

Big Data Framework for Indian Green Coffee Export Demand Modeling and Descriptive Analysis using Nosql-MONGODB

Saivijayalakshmi Janakiraman and N Ayyanathan

October 26, 2021

Big Data Framework for Indian Green Coffee Export Demand Modeling and Descriptive Analysis using Nosql-MONGODB

Author: Saivijayalakshmi Janakiraman Research Scholar B.S.Abdur Rahman Crescent Institute of Science and Technology, Chennai

Dr N.Ayyanathan, Associate Professor,Department of Computer Applications, B.S.Abdur Rahman Crescent Institute of Science and Technology, Chennai

Abstract:

Big data is creating many opportunities for different and diverse fields to achieve deeper and faster insights that can enhance the decision making. This paper addresses, analyzing and forecasting the export trend in Indian Green Coffee using various descriptive analytics and visualization techniques. With the rapid growth of technologies, large amount of data is produced from different sources that can either be structured or unstructured. Such type of data is very difficult to process and manage as it contains millions of records of information sourced from social media, web sales, audios, videos, images etc. Timely analysis of this data is a key factor for success in many business and service domains. Calculating descriptive statistics is a vital step while conducting research and should always be done before making inferential statistical comparisons. For this purpose MongoDB (Nosql) is made extensive use of in this paper.

Keywords: Bigdata, Indian Green Coffee, Descriptive Analytics and MongoDB.

1 Introduction

India is the third-largest producer and exporter of coffee in Asia, and the sixth-largest producer and fifth-largest exporter of coffee in the world. Though we have the potential to grow & remain as leading exporter, still we could not gain this position and even if we do, its not sustainable. While we attempted to find out the reason, the requirement of analysis of huge loads of data consisting of numerous variables as well as its variability depending on seasons, places, people, economy, pricing, logistics etc literally renders inability to analyze & conclude. An efficient supply chain results in better wealth management, more revenue generation, improved customer satisfaction and increased organizational productivity. In order to ripe the fruits of supply chain management; it needs to be managed properly [25]. Big data analytics is an integrated form of data analytics and web analytics for big data[29]. Big data and its emerging technologies including big data analytics have been not only making big changes in the way the e-commerce and e-services operate but also making traditional data analytics and business analytics bring new big opportunities for academia and enterprises [29]

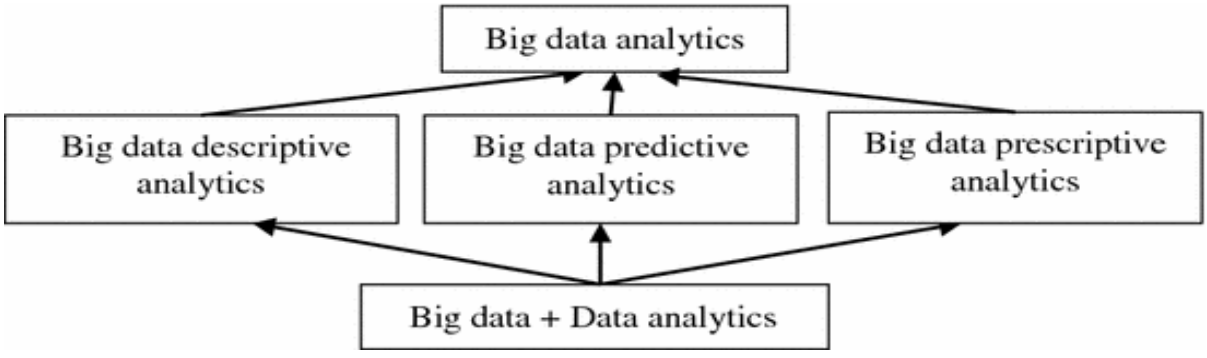

Figure 1: An Outlook of the Big Data Analytics

This paper is divided into 5 sections. Section 2 will provide an overview of a Literature Review. Section 3 will briefly discuss about the data description. Section 4 will review and discuss about the implementation of No-Sql MONGODB and techniques for Descriptive analytics across various processes of Big Data. The last section will conclude this paper.

1.1 Problem Definition:

As such, how to develop a systematic collection of data which are relevant, quantifiable, sort able and covers all variables was the 1st point to be catered, then how to relate between these data, what type of analytical tools to be used, what kind of graphs & pictorial representations are effective, how inferences can be drawn to be worked out.

Once we find the right process & tools, we should go about finding what points to be answered, leading to the conclusion. Our point of conclusion should be oriented to find out what are the grey areas we need to focus and improve upon to gain the position of ever leading Coffee exporter.

2 Literature Review

Business examination alludes to the broad utilization of information, procured by different sources, factual and quantitative investigation, illustrative and prescient models, and truth based administration to drive choices and activities to legitimate partners (Davenport and Harris, 2007;Soltanpoor&Sellis, 2016)[21].To do this, business analytics utilizes methods from the data science, operational research, machine learning and information systems fields (Mortenson, Doherty, & Robinson, 2015)[15]. In this sense, business analytics deal not only with descriptive models but also with models capable of providing meaningful insights and supporting decisions about business performance. To this end, business investigation has advanced past a basic crude information examination on huge datasets with an intend to give associations an upper hand (Mikalef, Pappas,Krogstie, and Giannakos, 2018;Vidgen, Shaw, and Grant, 2017)[14].

Business analytics is categorized to three main stages characterized by different levels of difficulty, value, and intelligence (i) descriptive analytics, answering the questions "What has happened?", "Why did it happen?", but also "What is happening now?" (mainly in a streaming context); (ii) predictive analytics, answering the questions "What will happen?" and "Why will it happen?" in the future; (iii) prescriptive analytics, answering the questions "What should I do?" and "Why should I do it?". (Akerkar, 2013; Krumeich, Werth, & Loos, 2016; Šikšnys& Pedersen, 2016)[16,17,18]

Soltanpoor&Sellis, stated that it helps organizations to grasp the reasons of the events that happened in the past and understand relationships among different kinds of data [9]. However, similarly to other research works we consider it as part of descriptive analytics. The reason for this is to ensure consistency among the three stages of analytics so that each one answers the questions "What?" and "Why?" (Krumeich et al., 2016; Šikšnys& Pedersen, 2016)[17,18],.

Big Data Analytics(BDA) has been applied in all stages of supply chains, including procurement, warehousing, logistics/transportation, manufacturing, and sales management. BDA consists of descriptive analytics, predictive analytics, and prescriptive analytics. Descriptive analytics is defined as describing and categorizing what happened in the past. Predictive analytics are used to predict future events and discover predictive patterns within data by using mathematical algorithms such as data mining, web mining, and text mining. Prescriptive analytics, apply data and mathematical algorithms for decision-making. Multi-criteria decision-making, optimization, and simulation are among the prescriptive analytical tools that help to improve the accuracy of forecasting [4]

The fundamentals of big data analytics consists of mathematics, statistics, engineering, human interface, computer science and information technology [28, 29]. There is a growing attention to analysis of consumption behavior and preferences using forecasts obtained from customer data and transaction records in order to manage

product supply chains (SC) accordingly [13,14]. Yang and Sutrisno applied and compared regression analysis and neural network techniques to derive demand forecasts for perishable goods. They concluded that accurate daily forecasts are achievable with knowledge of sales numbers in the first few hours of the day using either of the above methods [7]

Supply chain management (SCM) focuses on flow of goods, services, and information from points of origin to customers through a chain of entities and activities that are connected to one another [8]. As Hofmann and Rutschmann indicated in their literature review, the key questions to answer are why, what and how big data analytics/machine-learning algorithms could enhance forecast accuracy in comparison to conventional statistical forecasting approaches [5].

Mekuria T., Neuhoff D. and Köpke U. T. (2004) in their study entitled " The status of Coffee Production and The Potential for Organic Conversion in Ethiopia", highlights in an international conference that the collapse of world coffee prices is contributing to a socio – economic decline affecting an estimate of 125 million people world – wide. The conclusion was drawn that Ethiopia has the potentials to produce certified organic high quality coffee due to the favorable growing conditions and the high quality coffee due to favorable growing conditions and the high diversity of genetic resources in coffee Arabica[13].

Dr Gopalsamy and M.ArulKumar stated in their paper that India's export performance of coffee with respect to USA till 2018 was indicating positive performance in terms of quantity and it was expected to increase. The increasing and decreasing value on both quantity and value denotes the quality and quantity of demand imposed by USA towards India[33].

Their study emphasizes big data analytics related to unstructured data that form at least 95% of the big data. They have also reviewed analytics technique for audio, video, text, and social media data and they propose and invent new tools and techniques for predictive analytics for structured data, big data are noisy, unreadable, and interrelated so there is a need to develop a new statistical technique for the data provided by social media text, audio and video. [11,12].

2.1 Review Summary

SNO	AUTHOR	TITLE/YEAR OF PUBLICATION	METHODOLOGY	INFERENCE
1	Punam K, Pamula R, Jain PK. GUCON	A two-level statistical model for big mart sales prediction. Conference: 2018 International Conference on Computing, Power and Communication Technologies (GUCON)	KNN,Support Vector Regression,Two level Statistical Model	Two level model outperformed the single model techiniques

2	Burney SMA, Ali SM, Burney S.	A survey of soft computing applications for decision making in supply chain management. ICETSS 2017,2018	Fuzzy Logic,Artifical Neural Network,Genetic Algorithm	All algorithm widely used for Data Mining in various areas within Supply Chain Management.
3	Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding	Data Mining with Big Data. IEEE transactions on Knowledge and Data Engineering, vol. 26, no. 1, January 2014	Mining from Sparse, Uncertain, and Incomplete Data Spare,HACE THEOREM	Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors
4	R. Soltanpoor, T. Sellis M.A. Cheema, W. Zhang, L. Chang (Eds.)	Prescriptive analytics for big data. 27th Australasian database conference: ADC 2016. Databases theory and applications, LNCS, Vol. 9877, Springer International Publishing, Sydney, NSW (2016)	ARIMA,Logistic Regression	decision making and process effectiveness by helping analysts get closer to tying outcomes to specific situations
5	N Ayyanathan, A Kannammal	Combined forecasting and cognitive Decision Support System for Indian green coffee supply chain predictive analytics IEEE 2015 International Conference on Cognitive Computing and Information Processing(CCIP)	ARIMA, Support Vector Machine	The LS-SVM predicted values and conventional SVM predicted values can be combined as per the research design to generate the new tracking signal spectrum and measure the accuracy

2.2 Research Gap

Upon my review of many research papers/journals, I found "ARIMA" model has been widely used for statistical analysis. I preferred & used "MONGODB" – a Big Data analytical tool which is highly suitable to analyse data that vary or change frequently or that are semi or unstructured for analysization and visualization.

3 Data Description

Last ten years data i.e from 2009 to 2020 has been considered for analyzing the exporting trend of coffee and this was collected from Coffee Board of India, Bangaluru. Indian Green Coffee export to the forty countries and the research dataset for the period 2013-2021 was collected from published sources of the Indian Coffee Board web portal [39] . Many countries are importing Indian green coffee every year through Coffee Board of India. Demand quantity in metric tonne is the main indicator of the total supply chain network[1].

Data set contains various fields like Country-wise - Coffee exports quantity, % to total exports quantity, Unit Rate etc. I have collections which contains all the structured as well as unstructured data using the above dataset. I connected with "MONGODB" Atlas and therein, I have created a cluster. Thereafter I imported my data into that cluster and visualized using charts that is available in Mongodb Atlas.

4 Methodology

Upon my review of many research papers/journals, I found "ARIMA" model has been widely used for statistical analysis. I preferred & used "MONGODB"Nosql – a Big Data analytical tool which is highly suitable to analyse data that vary or change frequently or that are semi or unstructured for analysization and visualization.

MONGODB has various tools and I used "Compass and Atlas". "Compass" interacts with our data with full CRUD functionality. "Atlas" is a global cloud data base service which I made use for visualization of data. I made comparisons of the data set having the explanatory variables like quantity, % to total exports quantity and Unit price.

For statistical, comparisions, I used "Python", "MONGODB" and MS Excel. Heatmap is a component of Python which pictorially represents variables in terms of distinct colours. I used it to represent the country-wise leading exports duly taking into account of many variables.

Clear examination includes depiction of information as far as frequencies, extents, mean, middle, quartiles, standard deviation, between quartiles range and so forth. Estimation of these insights relies upon kind of factors either to be subjective or quantitative. Qualitative factors are all out, portrayed and inferable - for example sex, financial status, torment level, treatment bunches and so on Then again quantitative factors are quantifiable, consistent and mathematical - for example age, stature, weight, torment score and so forth.

Plain or graphical show of results is an extraordinary resource that a creator can use to introduce muddled and enormous volumes of discoveries. The pie outline and bar graph are utilized to introduce extents and frequencies got for subjective information. However, pie diagram isn't much ideal as it addresses one variable in particular. Additionally, with enormous number of classes, the pie diagram portrayal turns out to be very obscure. Mainly, there are two types of analyses involved in statistical findings. One is descriptive, another is inferential. In descriptive statistics, researcher only describes the findings of the collected data.

Pearson' correlation coefficient (**PCC**), or the **bivariate correlation**, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.Visualization is generally easier to understand than reading tabular data, heatmaps are typically used to visualize correlation matrices. This matrix tells a lot about the relationships between the variables involved. I could find a correlation of 1.0 along **the diagonal of the matrix**. This is because each variable is highly and positively correlated with itself.

4.1 Experimental Setup

Initially need to install MongoDB tools and MongoDB Atlas. Go to the download page www.mongodb.com select MongoDB Community Server as well the operating system. Once installation got over we can go and run the mongodb using mongodb.exe. From there we can access our databases.

Create a database and collections and add data using CRUD operations. After adding data's we can manipulate the data's too.We can also import data (file type should be in CSV or JSON format). Connect MongoDB Compass with MongoDB Atlas by using connection string then we can use those collections in the Atlas where I have created charts and my dashboards.

5 Results and Discussion

From figure 2 I inferred an end that Quantity is expansions in decrease in Price.(from 2012 onwards).From fig 3. I prepared bar chart for the various destination based on the exporting quantity through the periods of years.

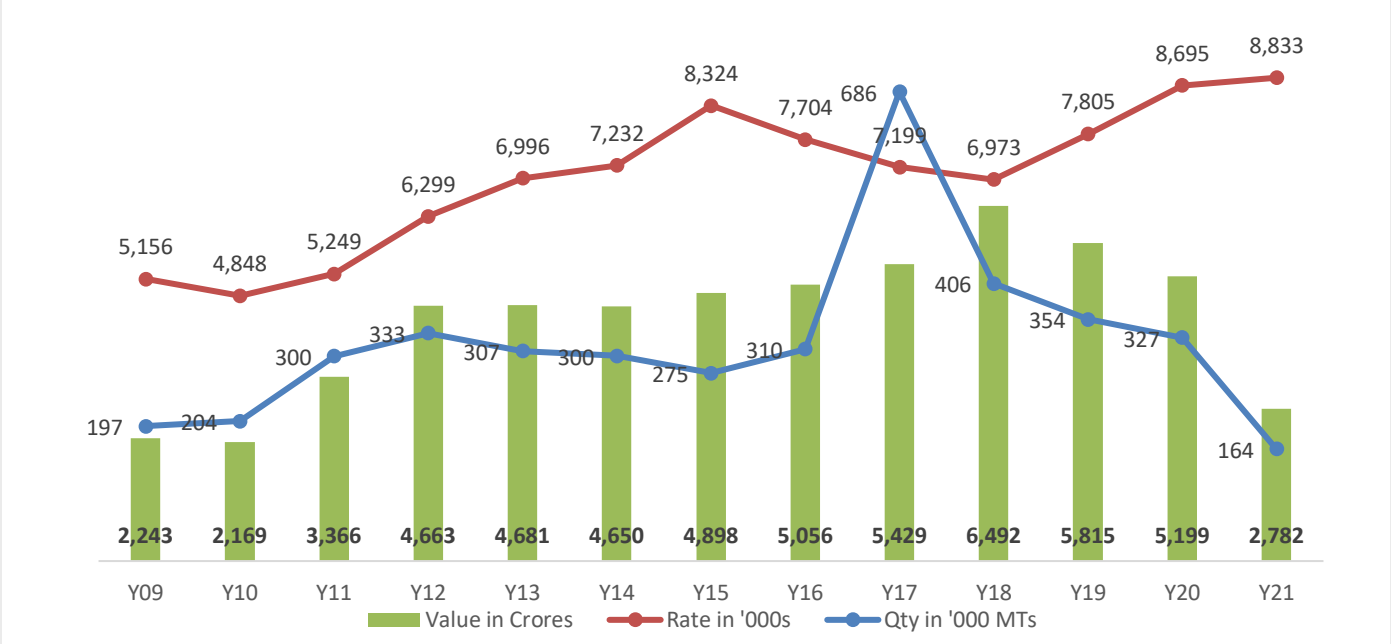

Figure2:Statistical Charts for Qty, Rate and Value of exports in different years

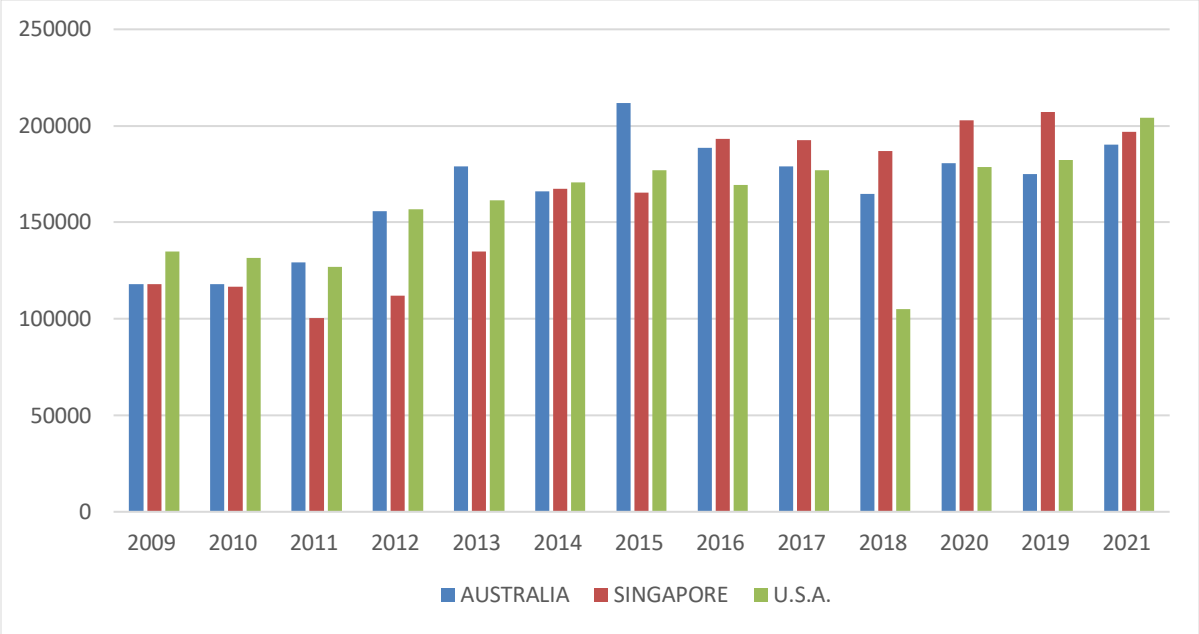

Figure3 Year-wise chart for the three countries based on Rate in Rs/M.Ton

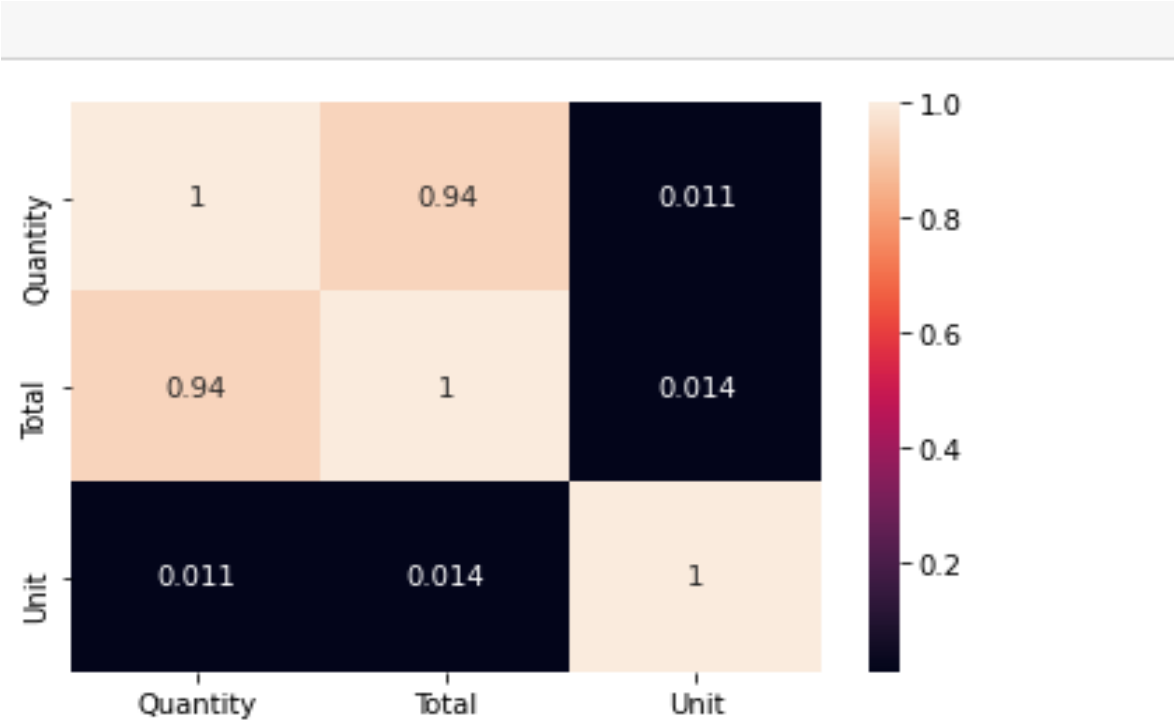

Figure 4:Correlation matrix shows a relationship between the explanatory variables.

6 Conclusion

The first step of data analysis is usually to describe the sample and then subgroups within the sample. Frequency distribution, mean, median, mode, range, and standard deviation are the most commonly used statistics for accomplishing this task. From the charts I can say that quantity increases with the reduction in price from 2012

onwards. In this paper, I presented an overview on Descriptive analytics tools and techniques that make it possible for analytics to build classifiers more effectively. I proposed to explore on deeper insights on factors influencing Coffee export demand, demand forecasting / predictions.

References

1. N Ayyanathan, A Kannammal : Combined forecasting and cognitive Decision Support System for Indian green coffee supply chain predictive analytics IEEE 2015 International Conference on Cognitive Computing and Information Processing(CCIP)
2. Prescriptive analytics: Literature review and research challenges :KaterinaLepeniotia, Alexandros Bousdekisa, DimitrisApostoloua,b, GregorisMentzasa,
3. Competing on analytics: The new science of winning T.H. Davenport and J.G. Harris Harvard Business Press (2007)
4. An Intelligent Approach to Demand Forecasting :Nimai Chand Das Adhikari∗ , NishanthDomakonda∗ , ChinmayaChandan∗ Gaurav Gupta∗ , RajatGarg∗ , S Teja∗ , Lalit Das∗ , Dr. AshutoshMisra†
5. Hofmann E, RutschmannE. :Big data analytics and demand forecasting in supply chains: a conceptual analysis. Int J Logist Manage. 2018;29(2):739–66. https://doi.org/10.1108/IJLM-04-2017-0088
6. Wang G, Gunasekaran A, Ngai EWT, Papadopoulos T. :Big data analytics in logistics and supply chain management: certain investigations for research and applications. Int J Prod Econ. 2016;176:98–110. https://doi.org/10.1016/J.IJPE.2016.03.014.
7. Yang CL, Sutrisno H. :Short-term sales forecast of perishable goods for franchise business. In: 2018 10th international conference on knowledge and smart technology: cybernetics in the next decades, KST 2018, p. 101–5; 2018. https://doi.org/10.1109/KST.2018.8426091.
8. Lu LX, Swaminathan JM. :Supply chain management. IntEncyclSocBehav Sci. 2015. https://doi.org/10.1016/B978-0-08-097086-8.73032-7.
9. R. Soltanpoor, T. Sellis:**Prescriptiveanalytics for big data** M.A. Cheema, W. Zhang, L. Chang (Eds.), 27th Australasian database conference: ADC 2016. Databases theory and applications, LNCS, Vol. 9877, Springer International Publishing, Sydney, NSW (2016), pp. 245-325
10. E. M. Frazzon, A. Albrecht, E. Israel, B. Hellingrath, A.-K. Cordes, and P. Saalmann: "Simulation model concept for evaluating spare parts supply chain planning methods," Industrial Informatics (INDIN), 2014 12th IEEE International Conference on, pp. 560 – 565, 2014.
11. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J InfManag. 2015;35(2):137–44.
12. Matz SC, NetzerO. :Using big data as a window into consumers' psychology. CurrOpinBehav Sci. 2017;18:7–12.
13. Mekuria T., Neuhoff D. and Köpke U. :The Status Of Coffee Production And The Potential For Organic Conversion In Ethiopia,Conference Paper · October 2004
14. Mikalef et al., 2018P. Mikalef, I. Pappas, J. Krogstie, M. Giannakos :Big data analytics capabilities: A systematic literature review and research agenda Information Systems and e-Business Management., 16 (3) (2018), pp. 547-578
15. M.J. Mortenson, N.F. Doherty, S. Robinson :Operational research from Taylorism to Terabytes: A research agenda for the analytics age European Journal of Operational Research, 241 (3) (2015), pp. 583-595
16. B. Akerkar (Ed.): Advanced data analytics for business Big data computing, CRC Press, Boca Raton (2013), pp. 373-397
17. J. Krumeich, D. Werth, P. Loos :Prescriptive control of business processes Business & Information Systems Engineering, 58 (4) (2016), pp. 261-280
18. L. Šikšnys, T.B. Pedersen :Prescriptive analytics L. Liu, M. Özsu (Eds.), Encyclopedia of database systems, Springer, New York, NY (2016)
19. Guo ZX, Wong WK, Li M. :A multivariate intelligent decision-making model for retail sales forecasting. Decis Support Syst. 2013;55(1):247–55. https://doi.org/10.1016/J.DSS.2013.01.026.

20. Wei J-T, Lee M-C, Chen H-K, Wu H-H. :Customer relationship management in the hairdressing industry: an application of data mining techniques. Expert Syst Appl. 2013;40(18):7513–8. https://doi.org/10.1016/J.ESWA.2013.07.053.
21. Davenport & Harris, 2007;Soltanpoor &Sellis, 2016:Competing on analytics: The new science of winning, Harvard Business Press
22. Y-T. Chang and S.-W. Sun:Arealtime interactive visualization system for knowledge transfer from social media in a big data, in lnforrmation, Communications and Signal Processing (ICICS) 2013 9th International Conference on. IEEE, 2013, pp. 1-5
23. Islek I, Oguducu SG. :A retail demand forecasting model based on data mining techniques. In: IEEE international symposium on industrial electronics; 2015, p. 55–60. https://doi.org/10.1109/ISIE.2015.7281443.
24. Kilimci ZH, Akyuz AO, Uysal M, Akyokus S, Uysal MO, Atak Bulbul B, Ekmis MA. :An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. Complexity. 2019;2019:1–15. https://doi.org/10.1155/2019/9067367.
25. S. A. Khan, "Importance of Measuring Supply Chain Management Performance," Industrial Engineering & Management, Dec. 2013
26. Punam K, Pamula R, Jain PK.: A two-level statistical model for big mart sales prediction. In: 2018 international conference on computing, power and communication technologies, GUCON 2018; 2019. https://doi.org/10.1109/ GUCON.2018.8675060.
27. Puspita PE, İnkaya T, AkanselM. :Clustering-based Sales Forecasting in a Forklift Distributor. In: UluslararasıMuhendislikArastirmaveGelistirmeDergisi, 1–17; 2019. https://doi.org/10.29137/umagd.473977.
28. Chen, C.P., Zhang, C.-Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. **275**, 314–347 (2014)
29. Sun, Z., Strang, K., Yearwood, J.: Analytics service oriented architecture for enterprise information systems. CONFENIS 2014, Hanoi, 4–6 December 2014. In: Proceedings of iiWAS2014, pp. 506-518. ACM Press (2014)
30. Burney SMA, Ali SM, Burney S.: A survey of soft computing applications for decision making in supply chain management. In: 2017 IEEE 3rd international conference on engineering technologies and social sciences, ICETSS 2017, 2018, p. 1–6. https://doi.org/10.1109/ICETSS.2017.8324158.
31. MajedKharfan and Vicky Wing Kei Chan, :Forecasting Seasonal Footwear Demand Using Machine Learning", *Publisher Massachusetts Institute of Technology*, 2018.
32. González Perea R, Camacho Poyato E, Montesinos P, Rodríguez Díaz JA:Optimisation of water demand forecasting by artifcial intelligence with short data sets. Biosyst Eng. 2019;177:59–66. https://doi.org/10.1016/J.BIOSYSTEMS ENG.2018.03.011.
33. M.M. Najafabadi, F. Villanustre, T.M. Khoshgoftaar , N. Seliya, R. :DOG DQG (0XKDUHPDJLF:Deep Learning Applications and Challenges in Big Data ´ Journal of Big Data, SpringerOpen, 2015.
34. J. CHEN, Y. CHEN et al., :Big data challenge: a data management perspective, Front. Comput. Sci, vol. 7, no. 2, pp. 157-164, 2013.
35. M. A. Beyer and D. Laney:The Importance of 'Big Data': A Definition, Gartner, 2012.
36. C. JI, Y. LI et al:,Big Data Processing : Big Challenges and Opportunities, Journal of Interconnection Networks, vol. 13, no. 3 & 4, 2012.
37. Coffee Export from India 1 M.Arul Kumar, 2 Dr S. Gopalsamy © May 2019 IJSDR | Volume 4, Issue 5
38. Matthew Herland, Taghi M Khoshgoftaar and Randall Wald: A Review of data mining using Big Data in Health Informatics in Journal of Big Data, Springer, vol. 1, no. 2, 2014.
39. www.indiacoffee.org–Indian Coffee Portal Available from <http//:www.indiacoffee.org>