Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 197 (2022) 685–692

Procedia Computer Science

www.elsevier.com/locate/procedia

Sixth Information Systems International Conference (ISICO 2021)

Diagnostic analysis for outlier detection in big data analytics

Fakhitah Ridzuan[a], Wan Mohd Nazmee Wan Zainon[a,b,*]

[a]*School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia*
[b]*Dental Simulation and Virtual Learning Research Excellence Consortium, Universiti Sains Malaysia, 13200 Penang, Malaysia*

Abstract

Recently, Big Data analytics has been one of the most emerging topics in the business field. Data is collected, processed and analyzed to gain useful insight for their organization. Big Data analytics has the potential to improve the quality of life and help to achieve Sustainable Development Goals (SDG). To ensure that SDG goals are achieved, we must utilize existing data to meet those targets and ensure accountability. However, data quality is often left out when dealing with data. Any types of errors presented in the dataset should be properly addressed to ensure the analysis provided is accurate and truthful. In this paper, we have addressed the concept of data quality diagnosis to identify the outlier presented in the dataset. The cause of the outlier is further discussed to identify potential improvements that can be done to the dataset. In addition, recommendations to improve the quality of data and data collection systems are provided.

Keywords: Big data; data quality; outlier; Sustainable Development Goals

1. Introduction

In recent years, Big Data and data analytics has become exceptionally valuable in many areas such as health, finance and retail. The benefits of Big Data analytics are not restricted to a specific type of industry since analytics has proven vital for organisations to stay on top its competitors. The growing enthusiasm for making a decision based on data creates the importance of accurate and precise prediction. Agenda for Sustainable Development Goals (SDG) demands a significant investment in time and resources because of its call to 'leave no one behind' [1]. To ensure this agenda is achieved, Big Data is seen as one of the important elements that can unravel the disparities in society that

* Corresponding author. Tel.: +604-653463; fax: +604-6573335.
E-mail address: fakhitah.ridzuan@gmail.com

10.1016/j.procs.2021.12.189

were previously hidden. There has been some research conducted shown that 65 SDG indicators could directly or indirectly benefit from Big Data sources [2]. The use of Big Data analytics has the potential to help organisations to realize the opportunities in supporting SDG, while at the same time targeting growth through a range of advantages in various business activities, including supply chain activities [3].

Food and Agriculture Organization of the United Nations (FAO) recently stated that the global food prices have jumped to 40% in May 2021, reaching their highest level in a decade [4]. Besides, the number of hungry people is increasing to 130 million, which take us to the levels of hunger that have not seen in half a century. This situation bothered us, on how we can do to overcome this. The current data ecosystem provides new opportunities as the organisations can leverage data and knowledge from others to improve their processes in collecting, processing, analysing and disseminating data [1]. However, the existence of dirty data inside the dataset is common and it is becoming an obstacle in providing accurate results and may result in misguided decisions [5]. Improving data quality is very important for data mining and data analysis to avoid losses, problems, and additional costs due to the poor quality of data.

Considering the vast amount of data available today, assessing and evaluating information reliability is even more critical. This process is time-consuming, and it is impossible for a human to identify reliable data from a large amount of information manually [6]. Experian reports that most of the companies lose about 12% of their sales due to wrong records, subsequently adding to reduced productivity, loss of resources, and significantly misused chances for the marketing of cross-channel [7]. Based on the survey conducted by Experian, approximately one-third of respondents think that they waste nearly 10% or more of their marketing budget because of the result from inaccurate data. 83% of the respondents stated that poor data quality has hurt their business objectives, while 66% reported that poor data quality has had a negative impact on their organisation in the last twelve months [8]. This demonstrates that the organisation needs to have an excellent quality of data in order to obtain more accurate and valuable results since they depend on the data like customer relationship management and supply chain management.

This article aims to identify the outlier issues inside the dataset and use the findings for the improvement of the data collections system. It is important to ensure that the data that will be analysed are in high quality and provide more truthful insight.

2. Literature review

2.1. Definition of big data

In general, Big Data refers to large and complex datasets that are often used for predictive analytics. It is a term that relates to massive, heterogeneous, and often unstructured digital data that are difficult to handle with traditional data management tools and techniques [9]. It is designed to gather, store, and manage the data through advanced analytic techniques and applications. The rapid growth of the data drives new business opportunities if the data is collected are handled correctly. However, data preparation is very challenging and time-consuming, as the volume and variety of data have been increasing in recent years [10]. The collected data need to be processed first to ensure its reliability before analytics can occur. This crucial phase will affect the result if unreliable data are used for the analysis. Thus, data cleansing is a critical process to make sure that the data is free from any type of errors. Big Data can generally be characterised into five main dimensions, called 5V's. Table 1 shows the explanation of Big Data criteria.

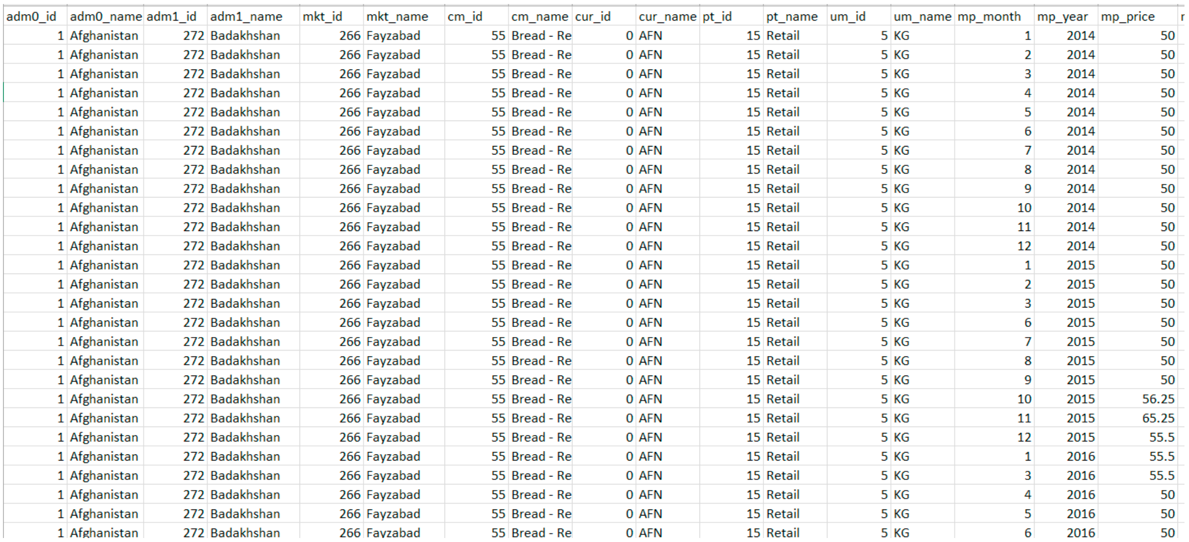
2.2. Data quality

'Garbage in garbage out' always happens in an organisation. Dirty data is collected, stored, processed, and analysed, but still provide the wrong decision, which leads to losses of a business organisation. Data quality is the primary concern faced by most organisations. This issue rises due to improper maintenance and will indirectly generate inconsistency in the database [11]. Data quality has become one of the most important data management issues as data from various sources in various formats are widely available [12].

Data quality is important to ensure the usability of the data. In this research, the Global Food Prices Dataset is obtained from Humanitarian Data Exchange [13] will be used for analysis purposes. This dataset is collected by WFP where food price information for 75 countries is included. The overview of the dataset is shown in Fig. 1. This dataset contains 1048576 records and 17 column listings which consist of the following attributes; Country Id, Country Name, State Id, State Name, Market Id, Market Name, Food Id, Food Name, Currency Id, Currency Name, Type Id, Type Name, UnitMetric Id, UnitMetric Name, Month, Year, price and Commodity Source. Price fluctuations and periodic prices hikes may affect people especially those coming from the low-income food-deficit country, as they increase food insecurity, hunger and numbers of people requiring food assistance [14].

Table 1. Explanation of big data criteria.

Big Data Criteria	Description
Volume	Big Data volume refers to the size of the data obtained, which is the most essential and distinctive feature of Big Data. Considering volume as the most effortless characteristic of Big Data to be defined, it also causes various challenges [15].
Velocity	Data velocity intends to measure how quickly the data are processed from the creation until the visualisation of the data. It can be created in real-time or near real-time. It is important as time-sensitive activities, such as fraud detection requires Big Data to analyse the pattern of real-time transactional data and find the correlation between multiple sources.
Variety	Variety refers to the different types of data that can be processed, which can consist of structured, semi-structured, or unstructured data. Text, images, videos, and audio are examples of the data stored in the database. The variety of data is expanding wildly as the handwriting of a person might be captured and stored as data.
Value	Data value refers to a measure of the usefulness of the collected data in decision-making [16]. It is closely related to volume and variety because it depends on the events or processes, such as stochastic, probabilistic, regular, or random [17].
Veracity	Veracity often relates to the trustworthiness and reliability of the data. It refers to how much amount of data can be trusted given the reliability of its source [16]. It is the ability of the data to help in the decision-making process by being appropriate, valuable, and of sufficient quality in the context in which it is analysed [18].



adm0_id	adm0_name	adm1_id	adm1_name	mkt_id	mkt_name	cm_id	cm_name	cur_id	cur_name	pt_id	pt_name	um_id	um_name	mp_month	mp_year	mp_price	r
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	1	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	2	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	3	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	4	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	5	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	6	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	7	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	8	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	9	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	10	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	11	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	12	2014	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	1	2015	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	2	2015	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	3	2015	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	6	2015	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	7	2015	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	8	2015	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	9	2015	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	10	2015	56.25	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	11	2015	65.25	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	12	2015	55.5	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	1	2016	55.5	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	3	2016	55.5	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	4	2016	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	5	2016	50	
1	Afghanistan	272	Badakhshan	266	Fayzabad	55	Bread - Re	0	AFN	15	Retail	5	KG	6	2016	50	

Fig. 1. Overview of the Global Food Price dataset.

This dataset is used to assess the level of sustainability of food choices available in the local market to achieve zero hunger. In addition, this dataset was also selected as it is publicly available and can be used as the benchmark for other researchers. World Food Programme (WFP) has been actively working with the national agriculture ministry to obtain the data of their market information system [14]. WFP has outlined eight objectives when collecting the data including the following:

1. Monitor food security

2. Provide information for cash-based-transfer
3. Initiate contact with the retailers
4. Monitor inflation trend
5. Assess the impact of the policy, program and project
6. Inform decision over time and location
7. Assist small farmers for agriculture planning and market decision
8. Increase information and transparency between producers, traders and consumers

Data quality represents the standard to which the information is fit for usage in the required business processes. It can be defined, measured, and managed through several data quality metrics, such as completeness, conformity, consistency, and accuracy. Data quality is not about how cleansed the data is, but it has become a critical issue because it involves operational processes and is perceived as the greatest challenge in data management [19].

Data quality issues arise due to poor data management in the organisation and this is happening at different levels; data source, generation level, and application level [20]. Thus, data quality must be controlled, particularly at the data source and generation level. In the data source, data quality is maintained through the data cleansing process. On the other hand, the generation level requires quality control when creating the dataset.

A review conducted by Corrales et al. [21] has shown that data quality issues for regression tasks, such as missing values, outliers, and redundancy have received greater attention from the research community. Meanwhile, since noise is characterised as a general consequence of data measurement errors, it has received less attention. To improve the quality of data, data cleansing is required [22] and it cannot be solved in one cycle of data cleansing only. It requires numerous iterations to verify that all the errors are being corrected, as some errors are only visible after the transformation [23]. Due to the abnormality of repaired data, the correctness of the data may not be achievable at the end of data cleansing and requires another cycle of the cleansing process [24].

Besides, data quality relies not only on its features but also on the data-driven business environment, including business processes and business users [25], which causes data quality to face several issues. Based on the above considerations, the basic question is how to evaluate the quality of the data to ensure that the data provided has high-level data quality.

2.3. Outlier

An outlier exhibits a significant deviation from the distributional behaviour of the remainders that make up a statistical sample [26]. It includes data values that deviate from the distribution of values in a column of a table. Characterizing, locating, and in some cases eliminating these outliers offers interesting insight about the data under scrutiny and reinforces the confidence that one may have in conclusions drawn from otherwise noisy datasets [27]. An outlier may not refer to the error inside the data, but it may potentially cause the data error. It can be divided into two types which are errors and unique samples [28]. Errors indicated the data has been mislabeled and could harm the model performance. On the other hand, a unique sample refers to the dataset which has a different structure from the majority of the dataset whereby it could help the model robustness. An outlier is not necessarily the wrong value as the comparison between the observation must be made before making the decision. For example, someone who is 200cm taller will be considered as the outlier in the general population, but when compared to the height of basketball players, the same person may not be considered an outlier.

There are six common techniques for outlier detection which are; Statistical-based methods, Distance-based methods, Density-based methods, Clustering-based methods, Deviation-based methods and Subspace-based methods [29]. Statistical methods are much more efficient as it is mathematically justified and the meaning behind the outlier can be easily revealed [30]. Besides, it is often presented in the graph and compact form which is easier to read. Therefore, in this paper, outlier detection methods using statistical-based methods will be used.

3. Data quality diagnosis

Data quality diagnosis was run on the dataset to understand the data and identify errors that appeared in the dataset. The main focus when running the quality diagnosis is findings the outlier as is important to predict the range values of the dataset. In this phase, the value will be evaluated by comparing them with the general distribution of the values inside the column. To detect numerical outliers, we need to compare the magnitude of the numerical data values. In this research, a histogram-based strategy is chosen as it is easy to understand the outlier from the graph plotted. It is the most popular non-parametric statistical technique to maintain a profile of data and it is simple to implement [30]. A histogram-based strategy builds a histogram distribution based on the frequency of data values in a particular data column. Eq. 1 shows the calculation for histogram-based strategy. The strategy $s_{\theta}tf$ marks data cells from the rare bins as data errors, i.e., data cells with a normalized term frequency smaller than a threshold $\theta_{tf} \in (0,1)$.

$$s_{\theta_{\mathrm{tf}}}(d[i,j]) = \begin{cases} 1, & \text{iff } \frac{TF(d[i,j])}{\sum_{i'=1}^{|d|} TF(d[i',j])} < \theta_{\mathrm{tf}} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $TF(d[i, j])$ is the term frequency of the data cell $d[i, j]$ inside the data column j.

To estimate the data distribution, mean, Q1, mean, median, and Q3, max will be used. If the number of zeros or minuses is dominant, then the data must be suspected to be skewed. If the number of outliers is large, strategies to eliminate the outliers are needed. If the value of the outlier is small, but the difference between the distribution with the outlier and the distribution without the outlier is very significant, thus it is necessary to remove the outlier in the dataset. Fig. 2 shows the visualisation of the price in the dataset with the outlier and without the outlier. It can be seen that the presence of the outlier provides a significant difference in the graph.

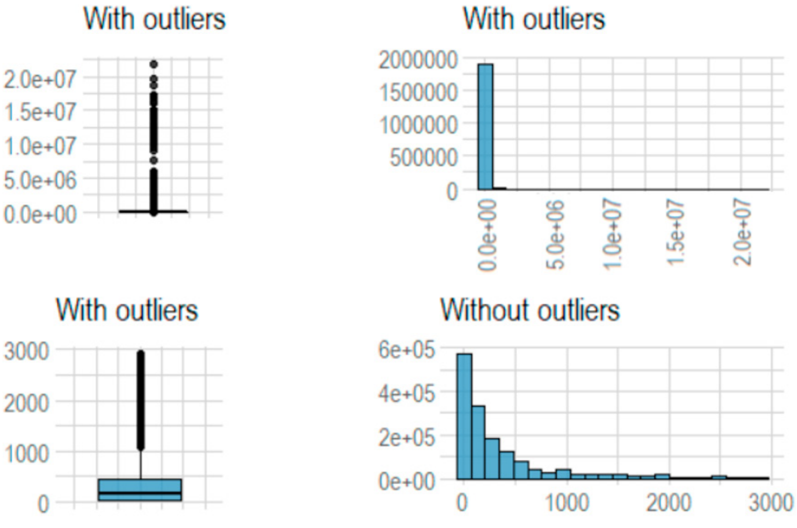


Fig. 2. Outliers in the Global Food Price dataset.

Numeric outliers are values that deviate significantly from the underlying distribution and it is common because of data entry errors and scale mismatch [31]. An outlier may not be dirty data and it cannot simply be eliminated. Generally, the presence of an outlier makes it difficult to analyse the data, but we should carefully look into the data and identify what are the contributing factors to the outlier. Since this research only focuses on error detection only, thus cleansing process will not be conducted.

4. Discussion

When dealing with the outlier, several things need to be examined before the cleansing process can be done. From the analysis, it can be seen that three main factors contribute to the outlier, which is; currency, year, prices, location and type of food. The dataset consists of the food price from a different country, and most probably the currency used will affect the outlier value. For example, 1 Algerian Dinar is equal to 0.64 Kyrgyzstani Som, looking at this value it is certain that the price in Kyrgyzstan will be higher than in Algeria. Thus, the dataset collection could further be improved by considering the use of standard currency to facilitate a comprehensive analysis.

Besides, the outlier may occur because of the difference in the year between each data. Food prices in increasing every year and this might cause the presence of outliers within the same country. Recently, a report by the United States Department of Agriculture has shown that food prices have risen at a rate of 2.0% to 3.0% per year on average [32]. Moreover, the quality of the food may also be improved which cause the price to rise steeply. This theory is backed up with evidence in [33] which the authors stated that the food prices would vary in one market or location, depends on its quality, freshness, location of the stall in the market, seller, and type of market. Therefore, a clear guideline on the collection of the data must be provided when obtaining the data to get cleaner data for the precise analysis.

Even if price series are not well correlated, it does not necessarily imply that the collected prices are wrong. Several factors create price differences besides data quality, for example, the import and export goods. Countries that export a certain type of food certainly promise cheaper food prices in their country compared to other countries. In addition, the type of district must be taken into account whether it is urban or rural because this position will affect the price of food. This was supported by a report by WFP [34], which stated that the outlier could also be correct data but an incredibly high price in a specific location.

Another factor that contributes to the outlier is the type of food. From the data, we can see the prices for retail and wholesale are combined, where the wholesale price is cheaper than retail. This will certainly contribute to the outlier. In addition, the price of a fish cannot be nearly the same as the price of a goat. When these prices are aggregated without being categorized or well-organized, it will cause outliers in the dataset. Therefore, it is important to reorganize the data that has been collected before performing the analysis.

To assess the quality of the dataset, still depending on the data stakeholders' skills and ability [35]. When an outlier is found, we cannot simply correct the error as we need feedback from the domain knowledge or expert. Only they know what are the specific criteria that can be identified as an outlier. Domain knowledge is an essential part of data cleansing and it is the human intelligence that accumulates within a specific practice or process [36]. The involvement of a domain expert in the data cleansing process is very important because detecting and correcting anomalies requires detailed domain knowledge [37]. However, a new issue emerges as the presence of domain knowledge is time and cost consuming. Consequently, automated techniques for obtaining and compiling rules to identify the correct outlier are needed to minimize the interaction with the domain knowledge.

Finally, it can be summarized that several factors may contribute to the presence of an outlier in the dataset. The cleansing process cannot be simply done automatically without any intervention from a domain expert as incorrect analysis will cause severe loss. To achieve SDG 2: Zero Hunger, monitoring of world food prices must be done to help in reducing the extreme food price volatility by ensuring proper market functioning of food commodities and their derivatives, as well as facilitating timely access to market information, including on food reserves.

The first phase in data cleansing is data analysis where errors and inconsistencies are identified in this phase [22]. To improve the quality of the dataset and ensure the sustainability of the data collection system, few things need to be considered:

i. Standardize the currency used - For the food price, it is difficult to see the price difference between each country if the currency used is not standardized. Data transformation is needed to ensure the data used for the analysis is presented in the same currency.

ii. Improve data storing process - The data storage process needs to be improved to facilitate the analysis that needs to be carried out. If all the information is combined in one table, the analysis will be more complex as we need to execute extra steps in the data cleansing process.

iii. Domain knowledge is needed – Before the cleansing process is executed, it is important to identify the rules, and these rules are coming from the expert. The original purpose of data collection needs to be identified to ensure that all information collected meets the purpose.

5. Conclusion

In this paper, we have provided an approach to identify the outliers in the datasets by using data diagnosis. Outlier must be handled properly in the cleansing process to make sure truthful insight can be provided. Furthermore, we have identified what are the factors that contribute to the outlier in the dataset. It can be seen that several factors greatly influence the presence of the outlier. The goal of this paper is to provide a way of improving the collection of the dataset to ensure all the errors are eliminated before the analysis process. Data is one of the important aspects to achieve SDG goals, thus it is important to have high-quality data.

References

[1] Cázarez-grageda, Karina, and Koffi Zougbede. (2019) "National SDG Review: Data Challenges and Opportunities", in *The Partnership in Statistics for Development in the 21st Century (PARIS21),* Paris, France.

[2] Halderen, Gemma Van, Irina Bernal, Tanja Sejersen, Ronald Jansen, Niels Ploug, and Maciej Truszczynski. (2021) "Big Data for the SDGs - Country Examples in Compiling SDG Indicators Using Non-Traditional Data Sources." *ESCAP Statistics Division Working Paper* No. SD/WP/12/January 2021.

[3] El-Haddadeh, Ramzi, Mohamad Osmani, Nitham Hindi, and Adam Fadlalla. (2021) "Value Creation for Realising the Sustainable Development Goals: Fostering Organisational Adoption of Big Data Analytics." *Journal of Business Research* **131**: 402–10.

[4] Smaller, Carin. (2021) "Food Prices Are Soaring, Hunger Is Rising: Here are three ways to stop another crisis." [Online]. Available: https://www.iisd.org/articles/food-prices-are-soaring-hunger-rising-here-are-three-ways-stop-another-crisis.

[5] Feng, Yuan. (2018) "Improve Data Quality by Using Dependencies and Regular Expressions." *Master's Thesis*, Mid Sweden University.

[6] Beretta, Valentina. (2018) "Data Veracity Assessment: Enhancing Truth Discovery Using a Priori Knowledge." *PhD Thesis*, IMT Mines Alès.

[7] M. Z. H, Jesmeen, Jakir Hossen, Md Shohel Sayeed, Chin Kuan Ho, Tawsif K., Md Armanur Rahman, and E. M. H. Arif. (2018) "A Survey on Cleaning Dirty Data Using Machine Learning Paradigm for Big Data Analytics." *Indonesian Journal of Electrical Engineering and Computer Science* **10(3)**: 1234–43.

[8] Heinrich, Bernd, Diana Hristova, Mathias Klier, Alexander Schiller, and Michael Szubartowicz. (2018) "Requirements for Data Quality Metrics." *Journal of Data and Information Quality* **9(2):** 1-32.

[9] Rodríguez-Mazahua, Lisbeth, Cristian Aarón Rodríguez-Enríquez, José Luis Sánchez-Cervantes, Jair Cervantes, Jorge Luis García-Alcaraz, and Giner Alor-Hernández. (2016) "A General Perspective of Big Data: Applications, Tools, Challenges and Trends." *Journal of Supercomputing* **72(8)**: 3073–3113.

[10] Tian, Yongchao. (2017) "Accelerating Data Preparation for Big Data Analytics." *PhD Thesis*, TELECOM ParisTech.

[11] Cohen, Bevin, David K. Vawdrey, Jianfang Liu, David Caplan, E. Yoko Furuya, Frederick W. Mis, and Elaine Larson. (2015) "Challenges Associated With Using Large Data Sets for Quality Assessment and Research in Clinical Settings." *Policy Politics and Nursing Practice* **16(3)**: 117–24.

[12] Côrte-Real, Nadine, Pedro Ruivo, and Tiago Oliveira. (2020) "Leveraging Internet of Things and Big Data Analytics Initiatives in European and American Firms: Is Data Quality a Way to Extract Business Value?" *Information and Management* **57(1)**.

[13] World Food Programme. (2021) "Global Food Prices Dataset 2021." [Online]. Available: https://data.humdata.org/dataset/wfp-food-prices.

[14] Food Security Information Network. (2015) "Review of Global Food Price Databases: Overlaps, Gaps and Opportunities to Improve Harmonization." [Online]. Available: https://reliefweb.int/report/world/review-global-food-price-databases-overlaps-gaps-and-opportunities-improve.

[15] L 'heureux, Alexanda, Katarina Grolinger, Hany F. Elyamany, and Miriam A. M. Capretz. (2017) "Machine Learning with Big Data: Challenges and Approaches." *IEEE Access* **5**:7776–7797.

[16] Swapnil, Walunj, Yadav Anil, and Sonu Gupta. (2016) "Big Data: Characteristics, Challenges and Data Mining." *IJCA Proceedings on International Conference on Advances in Information Technology and Management ICAIM* **2016(2)**: 25–29.

[17] Hadi, Hiba Jasim, Ammar Hameed Shnain, Sarah Hadishaheed, and Azizah bt Haji Ahmad. (2015) "Big Data and Five V'S Characteristics." *International Journal of Advances in Electronics and Computer Science* **2(1)**: 16–23.

[18] Crone, Robert. (2016) "Big Data Veracity Assessment". *PhD Thesis*, Delft University of Technology.

[19] Ehrlinger, Lisa, Elisa Rusz, and Wolfram Wöß. (2019) "A Survey of Data Quality Measurement and Monitoring Tools." CoRR abs/1907.08138.

[20] Taleb, Ikbal, Mohamed Adel Serhani, and Rachida Dssouli. (2018). "Big Data Quality: A Survey." in *2018 IEEE International Congress on Big Data*: 166–73.

[21] Corrales, David Camilo, Juan Carlos Corrales, and Agapito Ledezma. (2018) "How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning." *Symmetry* **10(4)**: 1–20.

[22] Ridzuan, Fakhitah, and Wan Mohd Nazmee Wan Zainon. (2019) "A Review on Data Cleansing Methods for Big Data." *Procedia Computer Science* **161**: 731–38.

[23] Guo, Aizhang, Xiuyuan Liu, and Tao Sun. (2018) "Research on Key Problems of Data Quality in Large Industrial Data Environment." in Proceedings of the 3rd International Conference on Robotics, Control and Automation - ICRCA '18, New York: *Association for Computing Machinery*. 245–48.

[24] Cheng, Hongju, Danyang Feng, Xiaobin Shi, and Chongcheng Chen. (2018) "Data Quality Analysis and Cleaning Strategy for Wireless Sensor Networks." *EURASIP Journal on Wireless Communications and Networking* **2018(61)**: 1-11.

[25] Cai, Li, and Yangyong Zhu. (2015) “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era.” *Data Science Journal* **14**: 1–10.

[26] Monhor, Davaadorjin and Shuzo Takemoto. (2005) “Understanding the Concept of Outlier and Its Relevance to The Assessment of Data Quality: Probabilistic Background Theory.” *Earth, Planets and Space* **57**:1009–1018.

[27] Pit--claudel, Clement, Zelda Mariet, Rachael Harding, and Sam Madden. (2016) “Outlier Detection in Heterogeneous Datasets Using Automatic Tuple Expansion.” *Computer Science and Artificial Intelligence Laboratory Technical Report*.

[28] Larson, Stefan, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael A. Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. (2019) “Outlier Detection for Improved Data Quality and Diversity in Dialog Systems.” *NAACL*: 517–527.

[29] Xu, Xiaodan, Huawen Liu, Li Li, and Minghai Yao. (2018) “A comparison of outlier detection techniques for high-dimensional data.” *International Journal of Computational Intelligence Systems* **11**:652–62.

[30] Zhang, Ji. (2013) “Advancements of Outlier Detection: A Survey.” *ICST Transactions on Scalable Information System* **13(1)**: 1-26.

[31] Wang, Pei. (2019) “Uni-Detect: A Unified Approach to Automated Error Detection in Tables.” in *SIGMOD '19: Proceedings of the 2019 International Conference on Management of Data*: 811–828.

[32] Amadeo, Kimberly. (2021) “Food Prices: 5 Causes of Increases, Trends, Forecast, Effect - 5 Causes of High Food Prices.” [Online]. Available: https://www.thebalance.com/why-are-food-prices-rising-causes-of-food-price-inflation-3306099.

[33] Donmez, Ayca, Gloria Solano-Hermosilla, Vladimir Bougay, Balaji Subbaraman, Robert M'barek, Abdoulaye Adam; Stephen BAHEMUKA; Oliver J. M. Chinganya, Vladimir Eskin; Koua Louis Kouakou; Charles Lufumpa; Rafik Mahjoubi; Ivo F. Njosa; Fabien Santini. (2017) “Using Web and Mobile Phone Technologies to Collect Food Market Prices in Africa, Approaching Real-Time Data and Use of Crowdsourcing, 2013 – 2016.” *Luxembourg, Publications Office of the European Union*.

[34] World Food Programme. (2020) “Introduction to Data Quality.” *In Fighting Hunger Worldwide*: 1–3.

[35] Jaya, M. Izham, Fatimah Sidi, Lilly Suriani Affendey, Marzanah A. Jabar, and Iskandar Ishak. (2017) “A Review of Data Quality Research in Achieving High Data Quality within Organization.” *Journal of Theoretical and Applied Information Technology* **95(12)**: 2647–57.

[36] Noraini, Abdullah, Saiful Adli Ismail, Siti Sophiayati, and Suriani Mohd Sam. (2015) “Data Quality in Big Data: A Review.” *International Journal of Advances in Soft Computing and its Applications* **7(Specialissue3)**: 16–27.

[37] Rahman, Fauziah Abdul, Rahimah Kassim, Zirawani Baharum, Helmi Adly Mohd Noor, and Norhaidah Abu Haris. (2019) “Data Cleaning in Knowledge Discovery Database-Data Mining (KDD-DM).” *International Journal of Engineering and Advanced Technology* **8(6 Special Issue 3)**: 2196–99.