An Assessment of Open Data Sets Completeness

By: Ali, A (Ali, Abdulrazzak)[1]; Emran, NA (Emran, Nurul A.)[2]; Asmai, SA (Asmai, Siti A.)[3]; Ismail, AR (Ismail, Amelia R.)[4]

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS
Volume: 10 **Issue:** 6 **Pages:** 557-562
Published: JUN 2019
Document Type: Article

Abstract

The rapid growth of open data sources is driven by free-of-charge contents and ease of accessibility. While it is convenient for public data consumers to use data sets extracted from open data sources, the decision to use these data sets should be based on data sets' quality. Several data quality dimensions such as completeness, accuracy, and timeliness are common requirements to make data fit for use. More importantly, in many cases, high-quality data sets are desirable in ensuring reliable outcomes of reports and analytics. Even though many open data sources provide data quality guidelines, the responsibility to ensure data of high quality requires commitment from data contributors. In this paper, an initial investigation on the quality of open data sets in terms of completeness dimension was conducted. In particular, the results of the missing values in 20 open data sets measurement were extracted from the open data sources. The analysis covered all the missing values representations which are not limited to nulls or blank spaces. The results exhibited a range of missing values ratios that indicated the level of the data sets completeness. The limited coverage of this analysis does not hinder understanding of the current level of data completeness of open data sets. The findings may motivate open data providers to design initiatives that will empower data quality policy and guidelines for data contributors. In addition, this analysis may assist public data users to decide on the acceptability of open data sets by applying the simple methods proposed in this paper or performing data cleaning actions to improve the completeness of the data sets concerned.

Keywords

Author Keywords: Data completeness; missing values; open data; open data sources; data collection
KeyWords Plus: MISSING DATA; IMPUTATION; OPTIMIZATION; VALUES

Author Information

Reprint Address: Ali, A (reprint author)

Univ Teknikal Malaysia Melaka, Fak Teknol Maklumat & Komunikasi, Ayer Keroh 76300, Melaka, Malaysia.

Addresses:

[1] Univ Teknikal Malaysia Melaka, Fak Teknol Maklumat & Komunikasi, Ayer Keroh 76300, Melaka, Malaysia

[2] Univ Teknikal Malaysia Melaka, Computat Intelligence Technol CIT Res Grp, Ayer Keroh 76300, Melaka, Malaysia

[3] Univ Teknikal Malaysia Melaka, Optimizat Modeling Anal Simulat & Scheduling Opti, Ayer Keroh 76300, Melaka, Malaysia

[4] Int Islamic Univ Malaysia, Dept Comp Sci, Kulliyyah ICT, POB 10, Kuala Lumpur 50728, Malaysia

Funding

Funding Agency	Grant Number
Universiti Teknikal Malaysia Melaka (UTeM)	

View funding text

Publisher

SCIENCE & INFORMATION SAI ORGANIZATION LTD, 19 BOLLING RD, BRADFORD, WEST YORKSHIRE, 00000, ENGLAND

Categories / Classification

Research Areas: Computer Science
Web of Science Categories: Computer Science, Theory & Methods

See more data fields

Cited References: 45

Showing 30 of 45 View All in Cited References page

(from Web of Science Core Collection)

1. **Estimating missing values of skylines in incomplete database** — **Times Cited:** 4
By: Alwan, A.A.; Ibrahim, H.; Udzir, N.I.; et al.
Second International Conference on Digital Enterprise and Information Systems (DEIS 2013) Pages: 220-9 Published: 2013

2. Title: [not available] — Times Cited: 3
By: Anderson, A.
Statistics for Big Data For Dummies. Published: 2015
Publisher: John Wiley & Sons

3. **A Simplified Systematic Literature Review: Improving Software Requirements Specification Quality with Boilerplates** — Times Cited: 1
By: Anuar, U.; Ahmad, S.; Emran, N. A.
9 MAL SOFTW ENG C MY Pages: 99-105 Published: 2016

4. **Missing data resilient decision-making for healthcare IoT through personalization: A case study on maternal health** — Times Cited: 4
By: Azimi, Iman; Pahikkala, Tapio; Rahmani, Amir M.; et al.
FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF ESCIENCE Volume: 96 Pages: 297-308 Published: JUL 2019

5. Title: [not available] — Times Cited: 4
By: Banasiewicz, A.
Marketing Database Analytics - Transforming Data for Competitive Advantage Published: 2013
Publisher: Routledge

6. **How can I deal with missing data in my study?** — Times Cited: 328
By: Bennett, DA
AUSTRALIAN AND NEW ZEALAND JOURNAL OF PUBLIC HEALTH Volume: 25 Issue: 5 Pages: 464-469 Published: OCT 2001

7. Title: [not available] — Times Cited: 1
By: Bhatia, P.
Data Mining and Data Warehousing: Principles and Practical Techniques Published: 2019
Publisher: Cambridge University Press

8. **Query Optimization for Dynamic Imputation** — Times Cited: 5
By: Cambronero, Jose; Feser, John K.; Smith, Micah J.; et al.
PROCEEDINGS OF THE VLDB ENDOWMENT Volume: 10 Issue: 11 Pages: 1310-1321 Published: AUG 2017

9. **Searching Dimension Incomplete Databases** — Times Cited: 11
By: Cheng, Wei; Jin, Xiaoming; Sun, Jian-Tao; et al.
IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING Volume: 26 Issue: 3 Pages: 725-738 Published: MAR 2014

10. **Understanding relations (installment #7)** — Times Cited: 29
By: Codd, E.F.
Bulletin of ACM SIGMOD Volume: 7 Pages: 23-28 Published: 1975

11. **Principled missing data methods for researchers** — Times Cited: 399
By: Dong, Yiran; Peng, Chao-Ying Joanne
SPRINGERPLUS Volume: 2 Article Number: UNSP 222 Published: 2013

12. Title: [not available] — Times Cited: 1
By: Emran, N. A.
Definition And Analysis Of Population-Based Data Completeness Measurement Published: 2011
Ph. D. dissertation
Publisher: University of Manchester

13. **Model-driven component generation for families of completeness** — Times Cited: 1
By: Emran, N. A.; Embury, S. M.; Missier, P.
CTIT WORKSHOP P SERI Pages: 123-132 Published: 2008
Publisher: QDB/ MUD, Auckland, New Zealand

14. **Measuring population-based completeness for single nucleotide polymorphism (SNP) databases** — Times Cited: 3
By: Emran, NA; Embury, S; Missier, P.
Advanced approaches to intelligent information and database systems Pages: 173-182 Published: 2014
Publisher: Springer, New York

15. **Data accessibility model using QRCode for lifetime healthcare records** — Times Cited: 4
By: Emran, NA; Leza, FNM.
World Appi Sci J Volume: 30 Pages: 395-402 Published: 2014

16. **Measuring Data Completeness for Microbial Genomics Database** — Times Cited: 3
By: Emran, Nurul A.; Embury, Suzanne; Missier, Paolo; et al.
INTELLIGENT INFORMATION AND DATABASE SYSTEMS (ACIIDS 2013), PT I, Book Series: Lecture Notes in Computer Science Volume: 7802 Pages: 186-195 Published: 2013

17. **Storage Space Optimisation for Green Data Center** — Times Cited: 1

By: Emran, Nurul A.; Abdullah, Noraswaliza; Isa, Mohd Noor Mat

MALAYSIAN TECHNICAL UNIVERSITIES CONFERENCE ON ENGINEERING & TECHNOLOGY 2012 (MUCET 2012) Book Series: Procedia Engineering Volume: 53 Pages: 483-490 Published: 2013

18. **Data Completeness Measures** — Times Cited: 4

By: Emran, Nurul A.

PATTERN ANALYSIS, INTELLIGENT SECURITY AND THE INTERNET OF THINGS Book Series: Advances in Intelligent Systems and Computing Volume: 355 Pages: 117-130 Published: 2015

19. Title: [not available] — Times Cited: 5

By: Faraway, J. J.

Linear Models with R Published: 2016

Publisher: Chapman and Hall/ CRC

20. **Missing Data Analysis: Making It Work in the Real World** — Times Cited: 2,634

By: Graham, John W.

ANNUAL REVIEW OF PSYCHOLOGY Book Series: Annual Review of Psychology Volume: 60 Pages: 549-576 Published: 2009

21. **Explaining Query Answer Completeness and Correctness with Minimal Pattern Covers** — Times Cited: 1

By: Hannou, F.-Z.; Amann, B.; Baazizi, M.-A.

VLDB Endowment Volume: 12 Pages: 14 Published: 2019

22. **Principles of Data Management and Presentation** — Times Cited: 2

By: Hoffmann, J. P.

PRINCIPLES OF DATA MANAGEMENT AND PRESENTATION Pages: 1-262 Published: 2017

23. **What to Do about Missing Values in Time-Series Cross-Section Data** — Times Cited: 361

By: Honaker, James; King, Gary

AMERICAN JOURNAL OF POLITICAL SCIENCE Volume: 54 Issue: 2 Pages: 561-581 Published: APR 2010

24. **INCOMPLETE INFORMATION IN RELATIONAL DATABASES** — Times Cited: 383

By: IMIELINSKI, T; LIPSKI, W

JOURNAL OF THE ACM Volume: 31 Issue: 4 Pages: 761-791 Published: 1984

25. **Evaluating Performance of Missing Data Imputation Methods in IRT Analyses** — Times Cited: 2

By: Kalkan, Omur Kaya; Kara, Yusuf; Kelecioglu, Hulya

INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION Volume: 5 Issue: 3 Pages: 403-416 Published: 2018

26. **The prevention and handling of the missing data** (View record in KCI-Korean Journal Database) — Times Cited: 107

By: Kang, Hyun

KOREAN JOURNAL OF ANESTHESIOLOGY Volume: 64 Issue: 5 Pages: 402-406 Published: MAY 2013

27. Title: [not available] — Times Cited: 309

By: Kantardzic, M.

Data mining: Concepts, models, methods, and algorithms Published: 2011

Publisher: Wiley, New York

28. **Numerical calculation and experiment of a heaving-buoy wave energy converter with a latching control** — Times Cited: 3

By: Kim, Jeongrok; Cho, Il-Hyoung; Kim, Moo-Hyun

OCEAN SYSTEMS ENGINEERING-AN INTERNATIONAL JOURNAL Volume: 9 Issue: 1 Pages: 1-19 Published: MAR 2019

29. Title: [not available] — Times Cited: 27

By: Little, R.; Rubin, D.

Statistical Analysis with Missing Data Published: 2019

Publisher: John Wiley & Sons.

30. **Regression model approach to predict missing values in the Excel sheet databases** — Times Cited: 3

By: Mahesh, Z. Kumar; Manjula, R.

IJCSET Volume: 3 Issue: 4 Pages: 130-135 Published: 2012

Showing 30 of 45 **View All in Cited References page**