Quality Matters: Understanding the Impact of Incomplete Data on Visualization Recommendation

Rischan Mafrur[1 [0000−0003−4424−3736]], Mohamed A. Sharaf[2 [0000−0002−3405−5224]], and Guido Zuccon[1 [0000−0003−0271−5563]]

[1] The University of Queensland, Brisbane, Australia
{r.mafrur,g.zuccon}@uq.edu.au
[2] United Arab Emirates University, Al Ain, UAE
msharaf@uaeu.ac.ae

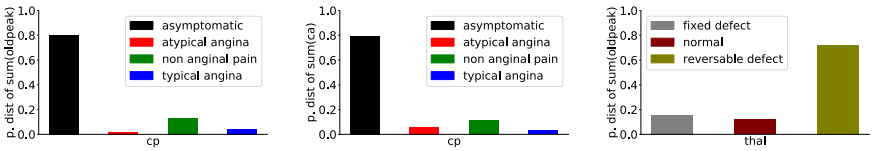
Abstract. Incomplete data is a crucial challenge to data exploration, analytics, and visualization recommendation. Incomplete data would distort the analysis and reduce the benefits of any data-driven approach leading to poor and misleading recommendations. Several data imputation methods have been introduced to handle the incomplete data challenge. However, it is well-known that those methods cannot fully solve the incomplete data problem, but they are rather a mitigating solution that allows for improving the quality of the results provided by the different analytics operating on incomplete data. Hence, in the absence of a robust and accurate solution for the incomplete data problem, it is important to study the impact of incomplete data on different visual analytics, and how those visual analytics are affected by the incomplete data problem. In this paper, we conduct a study to observe the interplay between incomplete data and recommended visual analytics, under a combination of different conditions including: (1) the distribution of incomplete data, (2) the adopted data imputation methods, (3) the types of insights revealed by recommended visualizations, and (4) the quality measures used for assessing the goodness of recommendations.

Keywords: Incomplete data, Visualization recommendation, Data exploration

1 Introduction

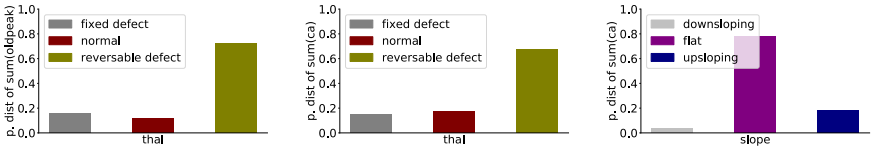
To support effective data exploration, there has been a growing interest in developing solutions that can automatically recommend data visualizations that reveal important data-driven insights. Several visual analytic tools have been introduced such as Tableau [9], Spotfire [8], Power BI [7]. The aim of those tools is to provide aesthetically high-quality visualizations that reveal interesting insights. Without any prior knowledge of the explored data, it is a challenging task for the analyst to manually select the combinations of attributes and measures that lead to interesting visualizations. Clearly, manually looking for insights in each visualization is a labor-intensive and time-consuming process. Such challenge motivated research efforts that focused on automatic recommendation of visualizations based on some metrics that capture the utility of recommended visualizations (e.g., [23], [22], [36], [17], [18], [35], [28], [15], [19]). However, all of those approaches operate under the assumption that the analyzed data is clean and overlook the data quality problems that might impair the recommendation process.

Data quality is a crucial challenge to data exploration and analytics. Poor data quality would distort the analysis and reduce the benefits of any data-driven approach. That



(a) top-1 (b) top-2 (c) top-3

Fig. 1: Top-k recommended visualizations obtained from complete heart disease dataset, $k = 3$



(a) top-1 (b) top-2 (c) top-3

Fig. 2: Top-k recommended visualizations obtained from incomplete heart disease dataset (20% missing values), $k = 3$, NaN values are ignored

is, *garbage in, garbage out (GIGO)* phenomenon. In real world settings, most datasets exhibit data quality problems, such as incomplete data, which in turn leads to incorrect analytical results (e.g., [26], [22]). This is true for *descriptive analytics*, in which incomplete data leads to incorrect results for aggregate and statistical queries [39]. It is also equally true for *predictive analytics*, where reduced accuracy in classification and prediction are common side effects of working with incomplete data (e.g., [10] [16]). Moreover, in the general context of recommendation systems, incomplete data has been shown to result in inaccurate rankings, which has the expected effect of producing poor and misleading recommendations [31].

Several *data imputation* methods have been introduced to handle the incomplete data challenge (e.g., [27], [24], [13]). However, it is well-known that those methods cannot fully solve the incomplete data problem, but they are rather a mitigating solution that allows for improving the quality of the results provided by the different analytics operating on incomplete data [10]. Hence, in the absence of a robust and accurate solution for the incomplete data problem, it remains especially important to study the impact of incomplete data on different visual analytics, and how those visual analytics are affected by the incomplete data problem. This has been the focus of several research studies, including assessing the impact of incomplete data on analytics that rely on aggregate and statistical queries [39], predictions and classifications (e.g., [10], [16]), or recommendation [31].

To the best of our knowledge, this work is the first to explore the impact of incomplete data on the quality of recommended visualizations. In particular, our focus in this work is to study the interplay between incomplete data and recommended visual analytics, under a combination of different conditions including: the distribution of incomplete data, the adopted data imputation methods, the types of insights revealed by those visualizations, and the quality measures used for assessing the goodness of recommendations.

To further illustrate the problems addressed in this work, consider the motivating example shown in Figures 1 and 2. Both figures show the recommended top-k visual insights from a heart disease dataset [4] under two different settings: (1) complete data

(Figure 1), versus (2) incomplete data, with 20% missing values (Figure 2). The detail about the figures is explained further in Figure 3a. In both settings, the top-k visual insights are generated using the deviation-based approach [36], where $k = 3$, and any missing cells (i.e., NaN values) are ignored.

Meanwhile, comparing Figures 1 and 2, we notice the following: 1) the recommendations from complete data (Figure 1) are significantly different from those on incomplete data (Figure 2); 2) the two sets of recommendations have only one visualization in common (i.e., visualization based on *sum oldpeak* vs. *thal*[3]); and 3) that one common visualization was ranked top-3 based on the complete data, whereas it is ranked top-1 based on the incomplete data!

Based on the example above, a user who is analyzing an incomplete data with 20% missing values, would obtain a top-k recommended visualizations that are significantly different from those obtained from a complete dataset, and in turn gaining "false" insights from the data. Since incomplete data is a prevailing problem that can only be slightly mitigated by data imputation methods, it becomes essential to evaluate and quantify its impact on the insights gained from visual data analytics approaches. That is precisely the goal of this work, in which our main contributions are summarized as follows:

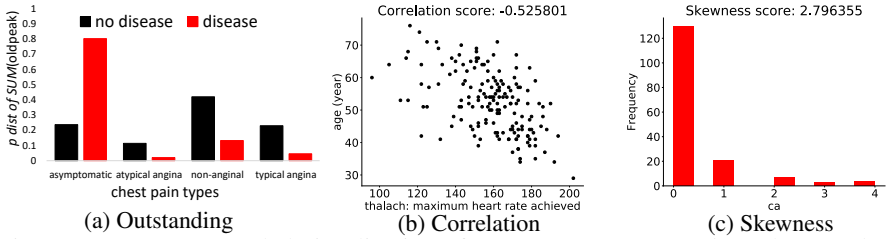
1. We study the different types of visual insights that are generally sought by data analysts in their data exploration workflows (Sec. 2).
2. We present three quality measures to quantify the impact of incomplete data on the quality of visualization recommendation (Sec. 3).
3. We conduct an extensive experimental evaluation on real datasets and present the impact of incomplete data on recommended visualizations with different data cleaning methods and different type of visual insights (Sec. 4).

2 Recommending Visual Insight

To recommend visual insight, we consider a multi-dimensional database D, which consists of a set of dimensional attributes $\mathbb{A}$ and a set of measure attributes $\mathbb{M}$. Also, let $\mathbb{F}$ be a set of possible aggregate functions over measure attributes. Hence, specifying different combinations of dimension and measure attributes along with various aggregate functions, generates a set of possible visualizations $\mathbb{V}$ over D. For instance, a possible visualization V_i is specified by a tuple $< A_i, M_i, F_i >$, where $A_i \in \mathbb{A}$, $M_i \in \mathbb{M}$, and $F_i \in \mathbb{F}$, and it can be formally defined as: Vi : `VISUALIZE bar (SELECT A, F(M) FROM D WHERE T GROUP BY A)`. Where `VISUALIZE` specifies the visualization type (i.e., bar chart), `SELECT` extracts the selected columns which can be dimensional attributes $A \in \mathbb{A}$ or measures $M \in \mathbb{M}$, T is the query predicate (e.g., disease = 'Yes'), and `GROUP BY` is used in collaboration with the `SELECT` statement to arrange identical data into groups.

Figure 1 shows the top-k recommended visual insights obtained from the complete heart disease dataset where $k = 3$. Figure 1a is obtained from V_i : `VISUALIZE bar (SELECT cp, SUM(oldpeak) FROM HeartDiseaseDB WHERE disease='Y' GROUP BY cp)`. However, obtaining this visualization V_i is only possible if the analyst knows exactly the parameters, which specify some aggregate visualizations that lead to those valuable visual insights (e.g., dimensional attributes, measures, aggregate

[3] thal: Thallium heart scan (normal, fixed defect, reversible defect)



(a) Outstanding (b) Correlation (c) Skewness

Fig. 3: Top-1 recommended visualization of (a) *aggregate/outstanding*, (b) *correlation* and (c) *skewness-based insight* from heart disease dataset where NaN values are ignored

functions, grouping attributes, etc.). Hence, it is time-consuming to iteratively create and refine visualizations to search for the ones that are useful and interesting.

Motivated by the need for efficient data analysis and exploration, several solutions for recommending visualizations have recently emerged (e.g., [36], [18], [29], [35], [28], [15], [14]). In such solutions, a large number of possible data visualizations $\mathbb{V}$ are generated and ranked according to some metrics that capture the *utility* of recommended visualizations. Towards this, the utility of each visualization V_i in $\mathbb{V}$ is calculated according to the type of insight, which is described next.

In this work, we study three types of visual insights: The first type is the *aggregate-based insight* which has been shown to be effective in recommending visualizations based on some metrics that capture the utility of a recommended visualizations (e.g., [36], [15], [35]). The second type is the *correlation-based insight*. This insight type is generally sought by data analysts looking for the attribute pairs with the highest correlations [14]. The third type is the *distribution-based insight* (e.g., *skewness* and *kurtosis*) (e.g., [32], [14]). In general, data analysts utilize distribution-based insight in order to find the dimensions that deviate from the normal distribution. Hence, by considering those insight types, we study insights based on single dimension (i.e., distribution-based insight), pairs of measures (i.e., correlation-based insight) and combination of dimensional attributes and aggregate functions of measures (i.e., aggregate-based insight). An example of those three types of visual insights can be seen in Figure 3. Given three types of the insights above, our problem definition as follows:

Definition 1. Recommending top-k visual insights: *Given a dataset D, insight type Y, the goal is to recommend a set top-k visual insight $S \subseteq \mathbb{V}$, where $|S| = k$, and $\mathbb{V}$ is the set of all possible generated visualizations from D, such that the overall utility $U(S)$ based on Y is maximized.*

Meanwhile, the utility of each visualization V_i is computed based on the type of insight shown by recommended visualizations, which are explained next.

2.1 Aggregate-based insight

In this paper, we address two types of aggregate-based insight: *outstanding* and *similarity* (e.g., [36], [34]). *Outstanding-based insight* recommends the most outstanding visualizations based on *deviation-based* approach (e.g., [36], [17], [29]). The deviation-based approach is able to provide analysts with interesting visualizations that highlight some of the particular trends of the analyzed datasets. The deviation-based approach compares an aggregate visualization generated from the selected subset dataset D_Q (i.e., target visualization $V_i(D_Q)$) to the same visualization if generated from a reference dataset D_R (i.e., reference visualization $V_i(D_R)$). To calculate the outstanding/deviation score, each target visualization $V_i(D_Q)$ is normalized into a *probability*

distribution $P[V_i(D_Q)]$ and similarly, each reference visualization into $P[V_i(D_R)]$. In particular, consider an aggregate visualization $V_i =< A, M, F >$. The result of that visualization can be represented as the set of tuples: $< (a_1, g_1), (a_j, g_j), ..., (a_t, g_t) >$, where t is the number of distinct values (i.e., groups) in attribute A, a_j is the j-th group in attribute A, and g_j is the aggregated value $F(M)$ for the group a_j. Hence, V_i is normalized by the sum of aggregate values $G = \sum_{j=1}^{t} g_j$, resulting in the probability distribution $P[V_i] =< \frac{g_1}{G}, \frac{g_2}{G}, ..., \frac{g_t}{G} >$. Finally, the utility score of V_i is measured in terms of the distance between $P[V_i(D_Q)]$ and $P[V_i(D_R)]$, and is simply defined as: $U(V_i) = dist\left(P\left[V_i(D_Q)\right], P\left[V_i(D_R)\right]\right)$

Figure 3a shows the top-1 recommended visualization of outstanding-based insight which is generated by [36] from heart disease dataset. The figure shows that an aggregate visualization based on *sum oldpeak* (i.e., pressure of the ST segment, where ST segment is an isoelectric section of the ECG) vs. *chest pain types* exhibits a large deviation between the target visualization (`disease`) and reference visualization (`no-disease`). That is, patients with a heart disease often suffer more from asymptomatic chest pains, in comparison to those without heart disease.

Meanwhile, *similarity-based insight* is the opposite to the outstanding-based insight. This insight type recommends the closest visualizations compared to the reference dataset [34].

2.2 Correlation-based insight

In the context of data exploration, data analysts generally derive insights from the data by iteratively computing and visualizing correlations looking for the attribute pairs with the highest correlations [14], either high positive or negative correlated [32]. Hence, the correlation-based insight recommends visualizations with the high correlated pair of measures. A visualization of correlation-based insight V_i is specified by a tuple $< B, C >$, where B and $C \subseteq \mathbb{M}$. The result of that visualization can be represented as the set of tuples: $< (b_1, c_1), (b_2, c_2), ..., (b_n, g_n) >$. Finally, the utility score of V_i is measured in terms of correlation coefficient of a tuple $< B, C >$. We use the Pearson correlation coefficient, which is formally defined as: $U(V_i) = \frac{\sum_{i=1}^{n}(b_i-\bar{b})(c_i-\bar{c})}{\sqrt{\sum_{i=1}^{n}(b_i-\bar{b})^2}\sqrt{\sum_{i=1}^{n}(c_i-\bar{c})^2}}$. Figure 3b shows top-1 recommended visualization V_i of correlation-based insight which is generated from the heart disease dataset, where V_i : `VISUALIZE scatter (SELECT thalach, age FROM HeartDiseaseDB WHERE disease='Y')`. The figure shows the high negative correlation of two measures (thalach: maximum heart rate achieved vs. age) where the correlation score is -0.53.

2.3 Distribution-based insight

Many classical statistical tests depend on normality assumptions [3]. Significant skewness and kurtosis clearly indicate that the data is not normaly distributed. Skewness is a measure of the lack of symmetry, while kurtosis is a measure of whether the data is heavy-tailed or light-tailed relative to a normal distribution. Generally, data analysts utilize values of skewness and kurtosis in order to find the attributes and measures that deviate from the normal distribution [12].

The distribution-based insight recommends the dimensional attributes or measures that most deviate from the normal distribution (e.g., [14], [12]). A visualization of distribution-based insight V_i is specified by a tuple $< E,$ `COUNT(E)`$>$. The utility score

for V_i is measured in terms of the third standardized moment μ_3 of V_i for the skewness-based insight and the fourth standardized moment μ_4 of V_i for the kurtosis-based insight. Hence, $U(V_i)$ for the skewness-based insight is $\frac{\mu_3}{\sigma^3}$, where $\mu_3 = \frac{\sum_{i=1}^{n}(e_i-\bar{e})^3}{n}$ and $\sigma = \frac{\sum_{i=1}^{n}(e_i-\bar{e})^2}{n}$. Meanwhile, $U(V_i)$ for the kurtosis-based insight is $\frac{\mu_4}{\sigma^4}$, where $\mu_4 = \frac{\sum_{i=1}^{n}(e_i-\bar{e})^4}{n}$ and $\sigma = \frac{\sum_{i=1}^{n}(e_i-\bar{e})^2}{n}$. In all cases, μ is the mean, σ is the standard deviation. Figure 3c shows the top-1 recommended visualization V_i of the skewness-based insight, where V_i : `VISUALIZE bar (SELECT ca, COUNT(ca) FROM HeartDiseaseDB WHERE disease='Y' GROUP BY ca)`. The figure shows ca is the dimension with the highest skewness score: $+2.8$, where ca is the number of major vessels colored by flourosopy.

3 Incomplete Data and Visualization Recommendation Quality

In this section, we first discuss the incomplete data problems (Sec. 3.1). Then, we introduce the quality measures used for assessing the quality of recommendations. (Sec. 3.2).

3.1 The incomplete Data Problem

Data quality is a crucial challenge to data exploration and analytics. Poor quality data would distort the analysis and reduce the benefits of any data-driven approach causing profound economic impact. Research has shown that the average cost of poor data on a business is 30% or more of its revenue [1]. The New York Times has also reported that analysts spend 50% - 80% of their time preparing dirty data before it can be used for data analytics [6]. Common examples of data quality challenges include multiple representations as a result of merging data from a variety of sources, incomplete data, anomalies, invalid, extreme, erroneous or duplicate values (e.g., [26], [22]).

In this paper, we focus on the incomplete data challenge. Incomplete data is common problem for data analytics (e.g., [26], [10], [16]). In *descriptive analytics*, incomplete data can lead to misleading conclusions such as wrong results for aggregate queries [39]. Meanwhile, in *predictive analytics*, incomplete data can introduce bias into a prediction or classification models (e.g., [16], [10]). Moreover, in the context of *recommendation systems*, incomplete data has been shown to result in inaccurate rankings, which has the expected effect of producing misleading recommendations [31].

Several data cleaning techniques have been introduced to overcome incomplete data issues include substituting missing data values by mean, median, or the most frequent value (e.g., [27], [24]), or using k-Nearest Neighbor [11], or association rules [38]. However, it is well-known that those imputation methods cannot fully solve the incomplete data problem. For instance, recent studies such as [10], [20] compared the performance of several imputation methods (e.g., median, linear regression) and showed the reduction of prediction and classification accuracy using those imputation methods.

Instead of proposing a new imputation method, this work investigates the impact of incomplete data on the quality of recommended visualizations. To the best of our knowledge, there is no prior work that focuses on that area. Existing work (e.g., [25], [33]) used sampling techniques to generate data visualizations and inspect the quality of the visualizations. However, our problem differs from those studies. Those studies focus on the quality of visualization while our work focuses on the quality of recommended visualizations. Another work is Profiler [22], which visualizes the data quality problems. This study also differs from ours. Profiler recommends visualizations that reveal data quality problems while our work recommends visualizations that reveal insights.

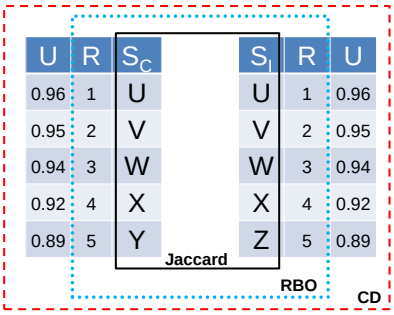


Fig. 4: A set of visualizations generated from complete data $S_C = (U, V, W, X, Y)$ and visualizations generated from incomplete data $S_I = (U, V, W, X, Z)$, R is ranking and U is utility score.

Recent data quality studies investigated the impact of incomplete data in predictive analytics (e.g., [10], [20]). Those studies compared the performance of various imputation methods on different supervised classifiers and explored the impact of incomplete data on the quality of classification and prediction models. Our problem differs from those studies in two ways. First, those studies focus on the impact of incomplete data in predictive analytics while our work is studying the impact of incomplete data in descriptive analytics. Second, the context of those studies are on general classification and prediction problems while our context is on visualization recommendation.

Toward investigating the impact of incomplete data on the quality of visualization recommendation, we introduce three measures for assessing the quality of recommendations, which explained next.

3.2 Quality of Recommended Visualizations

Recall from definition 1 that the goal of visualization recommendation is to recommend a set of top-k visualizations that reveal insights, in particular, as formulated in the definition 1, given a multi-dimensional dataset D, the set of top-k visualizations S is recommended. Let us consider D_I is the incomplete version of D. To facilitate the discussion, let us assume S_C is the set of top-k visualizations from the complete data, and it is equally to S. Moreover, S_I is the set of top-k visualizations from the complete data D. In order to understand the interplay between incomplete data and recommended visualizations, the top-k set obtained from an incomplete data S_I is compared to the top-k set obtained from the complete data S_C.

In this work, we utilize various metrics to assess the quality of the recommended visualizations in S_I compared to S_C. First, we utilize the *Jaccard distance* [30], which compares the composition of two sets as in Figure 4. The score of Jaccard distance is calculated by the number of visualizations in common, divided by the total number of visualizations. Accordingly, when applied to the set comparison, two sets with the same composition will have the same similarity score. However, in our work, the order of visualizations in the top-k set is essential. For instance, the top-1 visualization is more important than the top-10 visualization. Hence, we utilize the second metric, *Rank Biased Overlap (RBO)* [37], to consider the visualization ranking when assessing the quality of recommendations. As shown in Figure 4, RBO considers the composition of the two sets and their ranking, and it can be seen within the blue dotted line.

Finally, we have two metrics to evaluates our recommended visualizations. However, both metrics only compare the composition of the sets without considering the utility score of each visualization inside the set. Thus, we utilize the third metric called *Cumulative Distance (CD)* [21]. This metric captures both the utility score of each visualization $U(V_i)$ and visualization ranking. Figure 4 within the red dashed line illustrates the scope of the CD metric. The detail of those three metrics is explained next.

Jaccard distance Jaccard distance [30] is defined as the magnitude of the intersection divided by the magnitude of the union of the two sets, which is formally defined as: $Jaccard(S_I, S_C) = \frac{|S_I \bigcap S_C|}{|S_I \bigcup S_C|}$. This distance is bounded by 1. The value is between 0 for no similarity and 1 for identical sets. According to Figure 4, consider S_C = (U, V, W, X, Y) and S_I = (U, V, W, X, Z), Jaccard distance score of S_I to S_C is $\frac{4}{6} = 0.66$. The score is obtained from the number of intersection (i.e., four visualizations in common U, V, W, X) divided by the union (i.e., six visualizations in total U, V, W, X, Y, Z). This computation is based on the composition of both sets, the visualization ranking inside the set is not counted. For instance, if the visualizations in S_I is reversed (i.e., S_I = (Z, X, W, V, U)), the Jaccard distance score is still same 0.66.

Rank Biased Overlap (RBO) Since Jaccard distance is discounting the visualization order, we utilize the second metric called RBO [37]. RBO is a popular metric in Information Retrieval, which commonly used for the problem of comparing two ranked lists. RBO is compatible with item order and also compatible with the dis-jointness problem (i.e., an item is present only in one ranked list). In this work, we adopt RBO to quantify the quality of recommended visualizations in S_I compared to S_C.

To calculate RBO score, RBO determines the fraction of content overlapping at different depths. Consider at each depth d, the intersection of sets S_I and S_C to depth d is: $I_{S_I,S_C,d} = S_{I:d} \bigcap S_{C:d}$. The size of this intersection is the overlap of sets S_I and S_C to depth d, $X_{S_I,S_C,d} = |I_{S_I,S_C,d}|$ and the proportion of S_I and S_C that are overlapped at depth d is their agreement, $A'_{S_I,S_C,d} = \frac{X_{S_I,S_C,d}}{d}$. Hence, the RBO score of S_I and S_C is defined as: $RBO(S_I, S_C, p) = (1-p) \sum_{d=1}^{\infty} p^{d-1} * A'_{S_I,S_C,d}$. Similar to Jaccard, RBO has a range between 0 and 1, where 0 means disjoint, and 1 means identical. The parameter p models the user's persistence which is the probability of the user continuing to the next visualization. In particular, the smaller p, e.g., $p = 0$, only the top-ranked visualization is considered, and the RBO score is either zero or one. Meanwhile, if $p = 1$, the evaluation becomes arbitrarily deep due to the probability of deciding to stop is 0. The suggested p value is 0.95 or 0.97 [37]. In this work, we used $p = 0.95$, it means that the first 20 ranks have 86% of the weight of the evaluation.

Consider the example in Figure 4, using RBO the effectiveness score of S_I in comparison to S_C is 0.84 due to the both sets S_I, and S_C have only one different visualization on the tail. The Y is the last visualization in S_C, while the Z is the last visualization in S_I. However, if both sets have different on the head (i.e., top-1 visualization), the RBO score is 0.7. This example shows the visualization ranking is counted in RBO.

Cumulative Distance (CD) We utilize Cumulative Distance as our third metric. We adopt CD from DCG (Discounted Cumulative Gain) [21]. Similar to RBO, the DCG metric is generally used in Information Retrieval. This metric is a popular method for measuring the quality of search results. It assumes that highly relevant results are more valuable than marginally relevant results, and the top result is more important than the

tail. The DCG works by combining the degree of relevance and the rank of the search results in a coherent way. Meanwhile, the DGC score is unbounded. Hence, we can use the *normalized DCG (nDCG)*. The nDCG is defined as the actual DCG performance for a search query divided by the ideal DCG performance. To the best of our knowledge, this work is the first to use CD (i.e., mapped from nDCG) in the context of visualization recommendation. In our work, the degree of relevance of the visualization V_i is the utility score $U(V_i)$, where the utility score $U(V_i)$ is calculated according to the type of insight as explained in Sec 2. The CD score of S_I to S_C is defined as the DCG of S_I divided by DCG of S_C : $\text{CD} = \frac{\sum_{i=1, i \in S_I}^{n} \frac{1}{\log_2(i+1)} * U_i}{\sum_{i=1, i \in S}^{n} \frac{1}{\log_2(i+1)} * U_i}$, where U_i is utility score of each visualization from the complete dataset D.

Accordingly, Jaccard and RBO score from the example in Figure 4 are 0.66 and 0.84. Those scores indicate that both sets have quite a lot of differences. However, when we look at the CD score, it provides a different perspective. The score of the CD is 0.99. It is close to 1 (i.e., almost identical). That is because the utility score of Y and Z is precisely same ($U(Y) = U(Z) = 0.89$), which means the degree of importance of both visualizations (Y and Z) is the same.

4 Experimental Evaluation

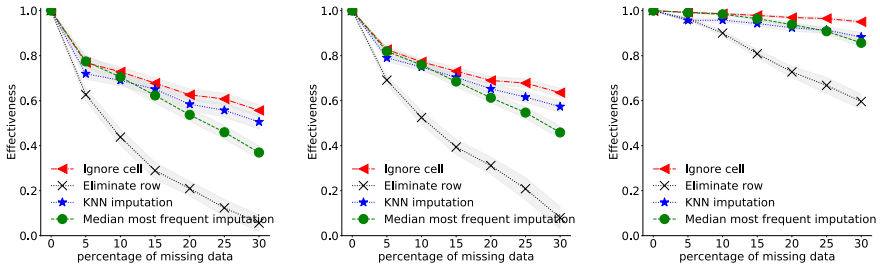
In this section, we first discuss our experimental testbed (Sec. 4.1). Then, we present and discuss our experimental evaluation. (Sec. 4.2).

4.1 Experimental Testbed

Data cleaning methods: In this work, we utilize and compare various well-known data cleaning methods, which are summarized as follows:

1. *Ignore cell*: The top-k visual insights are generated directly from the incomplete dataset by ignoring missing cells. In this approach, the process of handling incomplete data is on the cell level (e.g., [10], [20]).
2. *Eliminate row* : The process of handling incomplete data is on the row or tuple level. Particularly, a row that contains a missing cell is dropped. If the amount of missing cells is large, it may end up eliminating a huge amount of data [27].
3. *Impute cell*: In this approach, we utilize two common imputation techniques:
   (a) *Median and most frequent imputation*: This approach works by calculating the median of the non-missing values in a column and then replacing the missing values with the median within each column if the missing values are numerical data. Meanwhile, if the missing values are categorical data (strings or numerical representations), the missing values are imputed with the most frequent values within each column (e.g., [10], [20]).
   (b) *KNN imputation*: This approach imputes the missing data by finding the k closest neighbors to the observation with missing data and then imputing them based on the non-missing values in the neighbors.

Datasets: We conduct our experiments over the following datasets: (1) The Cleveland heart disease dataset is comprised of 8 dimensional attributes, 6 measures, and 299 tuples [4]. (2) The New York Airbnb dataset is comprised of 4 dimensional attributes, 4 measures, and 30249 tuples [5]. (3) The Diabetes 130 US hospital dataset consists of 14 dimensional attributes, 13 measures and 100 thousand tuples [2]. We conduct our



(a) Jaccard (b) RBO (c) CD

Fig. 5: Impact of data cleaning methods on effectiveness of outstanding-based insight using different data cleaning methods, $k = 10$

experiments over those three datasets, however, due to space limit, the Cleveland heart disease dataset is the default dataset for presenting the results in this paper.

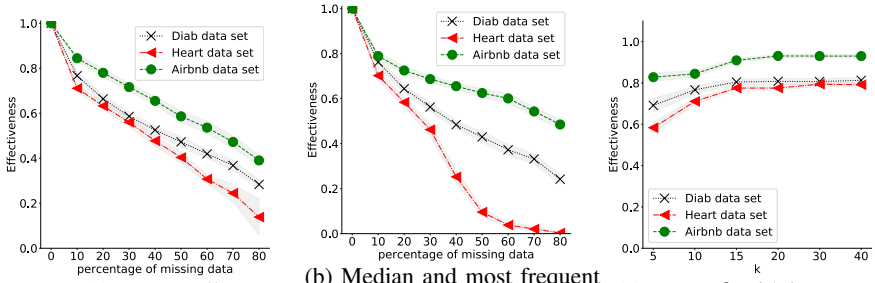
Incomplete data: We simulate missing data completely at random (MCAR) with different settings: (1) the distribution of missing values is on dimensional attributes $\mathbb{A}$, (2) the distribution of missing values is on dimensional measures $\mathbb{M}$, and (3) the distribution of missing values is on the whole data $\mathbb{A} + \mathbb{M}$. Recall from definition 1, in this experiment, we create an incomplete version data D_I from D. Then compare the top-k set S_I, which generated from the incomplete data D_I to the top-k set S_C, which generated from complete data D. In order to avoid bias, 100 versions of D_I with different random missing seed are generated. Finally, we repeat the experiments with different settings including: the percentage of missing values (i.e., 0% - 90%), the number of k, the type of insights, the data cleaning methods, and the quality measures used for assessing the quality of recommendation.

Default parameters: The default parameters used in our evaluation are $k = 10$, the percentage of missing data is 10%, the default of data cleaning method is *ignore cell*, the default dataset is Cleveland heart disease. The final result is the average from 100 versions of D_I and we present the results with confidence interval $CI = 0.95$.

Aggregate-based insight: In the case of aggregate-based insight, we use five aggregate functions (`COUNT, AVG, SUM, MIN and MAX`) where `COUNT` is only `COUNT(*)`. We use different query predicates T to understand the impact of input queries on the quality of recommendation with different percentages of missing values. For example, we want to compare an aggregate visualization generated from the selected subset dataset *chest pain types = 'typical angina'* to the visualization if generated from a reference dataset *chest pain types != 'typical angina'*. In this work, to study the impact of query predicate T on the quality of recommendation, we use three different queries for heart disease dataset: 1) q_1: *cp = typical angina* vs *cp != typical angina*; 2) q_2: *sex = Female* vs *sex = Male*; 3) q_3: *exang = exercise induced angina* vs *exang != exercise induced angina*.

4.2 Experimental Evaluation

In this section, we discuss our experiment results under a combination of different settings including: (1) the adopted data imputation methods, (2) the distribution of incomplete data, (3) the types of insights revealed by those visualizations, and (4) the quality measures used for assessing the quality of recommended visualizations.

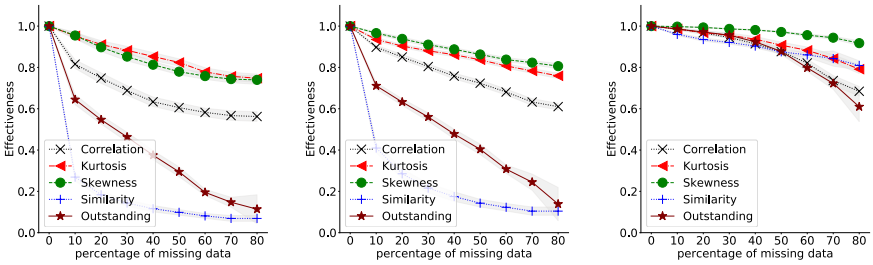


(a) Ignore cell (b) Median and most frequent imputation (c) Impact k with ignore cell

Fig. 6: Impact of data cleaning methods on effectiveness with different datasets - (a,b) *ignore cell vs. median and most frequent imputation*, (c) impact of k on effectiveness using *ignore cell* method

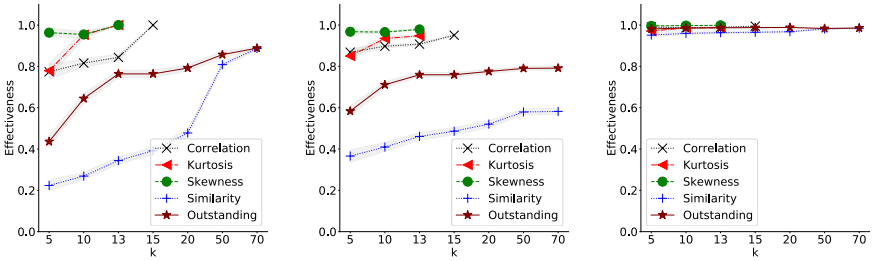
Impact of the data cleaning methods on effectiveness In this experiment, we analyze the effectiveness of data cleaning methods under different percentage of missing values and the quality measures (Jaccard, RBO, and CD). We compare four common data cleaning methods (e.g., *ignore cell, eliminate row, median and most frequent imputation*, and *KNN imputation*). Since the *eliminate row* method is included, the maximum percentage of missing values for this experiment is 30%. Moreover, the missing values are distributed on the whole data and the results of this experiments are generated based on the outstanding-based insight. As shown in Figure 5, the best data cleaning method is *ignore cell* and the worst is *eliminate row*. That is because that *eliminate row* leads to eliminate a huge amount of data. To the contrary, by ignoring missing cells without eliminating row, *ignore cell* outperforms other data cleaning methods. Meanwhile, in terms of imputation methods, *KNN imputation* has a better effectiveness than *Median and most frequent imputation* method. The result shows that the patterns are consistent for the three quality measures.

Impact of the data cleaning methods on different datasets In this experiment, we analyze the effectiveness of data cleaning methods under different datasets. We compare two data cleaning methods, which are *ignore cell* and *median and most frequent imputation* and the results of this experiments are generated based on the outstanding-based insight. The missing values are distributed on the whole data and maximum percentage of missing values for this experiment is 80%. As shown in Figure 6, overall, the pattern from three datasets are similar. In particular, in terms of the impact of missing values (Figures 6a and 6b), the effectiveness is decreasing when the number of missing values are increased. Moreover, in terms of the impact of k (Figure 6c), the effectiveness is increasing when k is increased. Meanwhile, if we compare Figures 6a and 6b, the effectiveness of *ignore cell* is better than *median and most frequent imputation*, especially for heart disease dataset. That is because the heart disease dataset has more dimensional attributes rather than measures. Imputing missing values on categorical data using *most frequent* method reduces the effectiveness. Further, the result of the Airbnb dataset is contrary to the result of the heart disease dataset. That is because the Airbnb dataset has more measures rather than dimensional attributes. The airbnb dataset consists of four dimensional attributes and four measures. However, since no incomplete data on predicate, the missing values are distributed on three dimensional attributes and four



(a) Jaccard (b) RBO (c) CD

Fig. 7: Impact of incomplete data on effectiveness of different insight types, $k = 10$



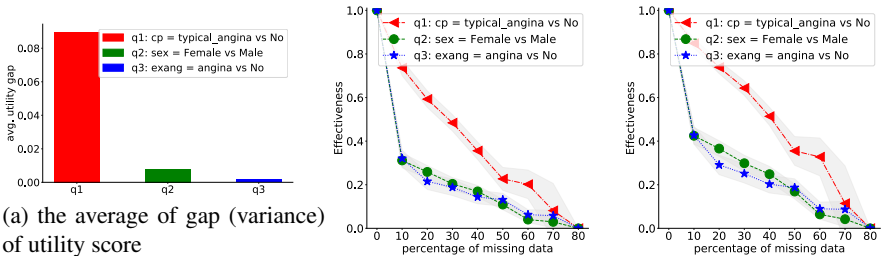
(a) Jaccard (b) RBO (c) CD

Fig. 8: Impact of k on effectiveness of different insight types, 10% missing values

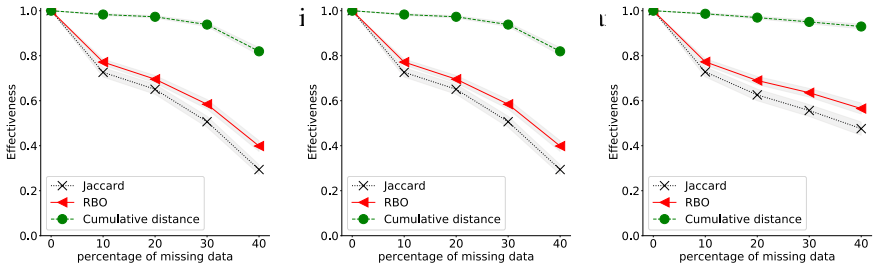
measures. Based on the results, we can conclude that *median and most frequent imputation* outperforms *ignore cell* if the data has more missing values on measures. **Impact of incomplete data on effectiveness** Figure 7 shows the impact of incomplete data on effectiveness under different types of insights. The figure shows that if the percentage of missing data is higher then it reduces the quality of visualization recommendation. The most resilient insight type to incomplete data is distribution-based insight (i.e., skewness, kurtosis), then the correlation-based insight, and the less resilient is aggregate-based insight. The skewness-based insight and kurtosis-based insight are specified by a single attribute or measure. Hence, losing a certain percentage of data will not change much of the data in each dimension. Meanwhile, the correlation-based insight is based on a pair of measures. Hence, the correlation-based insight less tolerance to the incomplete compared to the distribution-based insight. The aggregate-based insight is the most complex insight type. It is specified by the combination of dimensional attributes and the aggregate function of measures. Hence, the aggregate-based insight is the most sensitive to incomplete data, especially the similarity-based insight.

Impact of k on effectiveness As shown in Figure 8, the higher number of k results in the higher effectiveness due to the probability of the top-k set from the incomplete data having same content to the top-k set from the complete data is higher if the number of k is larger. For instance, Jaccard score is equal to 1, if $k = |\mathbb{V}|$, where the number of k equal to the number of candidate visualizations, however, it is only applies to Jaccard not to RBO and CD.

Impact of input queries on effectiveness Figure 9 shows the impact of predicate queries on the quality of visualization recommendation. Three different queries are used: 1) q_1: *cp = typical angina* vs *cp != typical angina*; 2) q_2: *sex = Female* vs *sex*

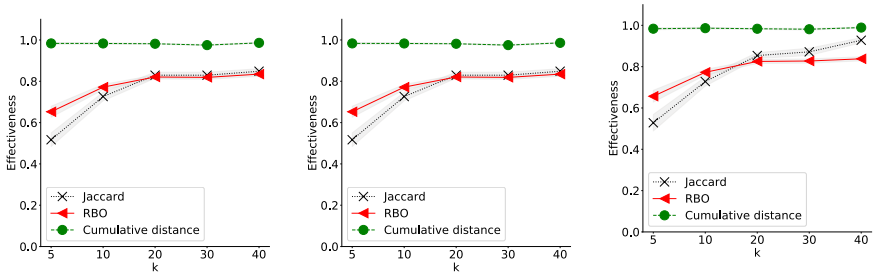


(a) the average of gap (variance) of utility score



(a) Impact dimensional attributes $\mathbb{A}$ on effectiveness (b) Impact measures $\mathbb{M}$ on effectiveness (c) Impact attributes and measures $\mathbb{A} + \mathbb{M}$ on effectiveness

Fig. 10: Impact of dimensional attributes, measures, and attributes + measures on effectiveness of outstanding-based insight, $k = 10$



(a) Impact k when the distribution of missing values on $\mathbb{A}$ (b) Impact k when the distribution of missing values on $\mathbb{M}$ (c) Impact k when the distribution of missing values on $\mathbb{A} + \mathbb{M}$

Fig. 11: Impact of k on effectiveness of outstanding-based insight 10% missing values

= *Male*; 3) q_3: *exang = exercise induced angina* vs *exang != exercise induced angina*. Figure 9a shows that $q1$ is more resilient to the incomplete data compared to other input queries (Figure 9b and 9c). The results show that if the input query generates top-k set that the variance among utility score of visualizations is very low, this low variance leads to more loss on effectiveness especially if the number of missing values is high.

Impact of dimensional attributes, measures, and attributes+measures on effectiveness Do the incomplete data on dimensional attributes have more impact rather than on measures? If so, when data analyst has a dataset with missing values on both dimensional attributes and measures, then she should give more attention to dimensional attributes rather than measures. Based on the experiment results, missing values on at-

tributes and measures have the same impact on the effectiveness. Figure 10 shows the impact of dimensional attributes, measures, and both on effectiveness with different percentage of missing values. The results are generated based on the heart disease dataset with the distribution of missing values on attributes and measures are equal. The results show that categorical and numerical data are equally important.

Impact of recommendation quality metrics on effectiveness using different number of k and different missing data distributions Figure 11 shows the impact of k on effectiveness if the incomplete data only on attributes, only on measures, and on both attributes and measures. As mentioned above, missing values on attributes and measures have the same impact on effectiveness (Figure 11a and 11b). The results also show how the performance of our three quality measures (i.e., Jaccard, RBO, and CD) under different number of k. Cumulative distance CD always performs above Jaccard and RBO. It is because of the default of percentage of missing values is quite small (10%). Meanwhile, there is an interesting pattern in Figure 11c, the figure shows that if the number of k is small (e.g., $5, 10$), Jaccard performs under RBO, however, when k is large (e.g., > 20), Jaccard performs above RBO and there is a crossover between both of them. Hence, the higher number of k results in the higher effectiveness in terms of Jaccard but not RBO. Jaccard score is equal to 1, if $k = |\mathbb{V}|$ where the number of k equal to the number of candidate visualizations. To the contrary, RBO has a different pattern, RBO score can be equal to 1 if visualizations inside the two top-k sets are in the same order, which is hard to be achieved. Hence, by increasing the number of k does not necessarily result in increased effectiveness in terms of RBO.

5 Conclusions

In this work, we investigate the interplay between incomplete data and recommended visual analytics under a combination of different conditions. This study lays the foundation for further exploring appropriate ways to deal with incomplete data and minimize the impact of incomplete data on visualization recommendation. We believe that this work can provide valuable insights for data analysts rather than blindly believing a recommendation result over low-quality data.

Acknowledgments: Rischan Mafrur is sponsored by the Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan / LPDP)(201706220111044). Dr Mohamed A. Sharaf is supported by UAE University Grant (G00003352). Dr Guido Zuccon is the recipient of an Australian Research Council DECRA Research Fellowship (DE180101579) and a Google Faculty Award.

References

1. Bad data cost, https://www.entrepreneur.com/article/332238
2. Diabetes 130 us hospitals 1999-2008, https://www.kaggle.com/brandao/diabetes
3. e-handbook of statistical methods, http://www.itl.nist.gov/div898/handbook/
4. Heart disease data set, https://archive.ics.uci.edu/ml/datasets/heart+Disease
5. Inside airbnb, http://insideairbnb.com/new-york-city/
6. Janitor work is key hurdle to insights, https://nyti.ms/1mZywng
7. Power bi, https://powerbi.microsoft.com/en-us/
8. Spotfire, https://www.tibco.com/products/tibco-spotfire/
9. Tableau, https://public.tableau.com/s/

10. Barata, A.P., et al.: Imputation methods outperform missing-indicator for data missing completely at random. In: ICDM (2019)
11. Batista, G.E.A.P.A., et al.: A study of k-nn as an imputation method. In: HIS (2002)
12. Bono, R., et al.: Bias, precision, and accuracy of skewness and kurtosis estimators for frequently used continuous distributions. SYMMAM **12**(1), 19 (2020)
13. Cambronero, J., et al.: Query optimization for dynamic imputation. PVLDB **10**(11), 1310–1321 (2017)
14. Demiralp, Ç., et al.: Foresight: Recommending visual insights. PVLDB **10**(12), 1937–1940 (2017)
15. Ding, R., et al.: Quickinsights: Quick and automatic discovery of insights from multi-dimensional data. In: SIGMOD (2019)
16. Ehrlinger, L., et al.: A daql to monitor data quality in machine learning applications. In: DEXA (2019)
17. Ehsan, H., et al.: Muve: Efficient multi-objective view recommendation for visual data exploration. In: ICDE (2016)
18. Ehsan, H., et al.: Efficient recommendation of aggregate data visualizations. TKDE **30**(2), 263–277 (2018)
19. Ehsan, H., et al.: Qurve: Query refinement for view recommendation in visual data exploration. In: ADBIS (2020)
20. Garciarena, U., et al.: An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Syst. Appl. **89**, 52–65 (2017)
21. Järvelin, K., et al.: Cumulated gain-based evaluation of IR techniques. TOIS **20**(4), 422–446 (2002)
22. Kandel, S., et al.: Profiler: integrated statistical analysis and visualization for data quality assessment. In: AVI (2012)
23. Key, A., et al.: Vizdeck: dashboards for visual analytics. In: SIGMOD (2012)
24. Khatri, H., et al.: QPIAD: query processing over incomplete autonomous databases. In: ICDE (2007)
25. Kim, A., et al.: Rapid sampling for visualizations with ordering guarantees. PVLDB **8**(5), 521–532 (2015)
26. Kim, W.Y., et al.: A taxonomy of dirty data. KDD **7**(1), 81–99 (2003)
27. Little, R.J.A., et al.: Statistical Analysis with Missing Data. John Wiley, USA (1986)
28. Luo, Y., et al.: Deepeye: Towards automatic data visualization. In: ICDE (2018)
29. Mafrur, R., et al.: Dive: Diversifying view recommendation for visual data exploration. In: CIKM (2018)
30. Manning, C.D., et al.: Introduction to information retrieval. Cambridge (2008)
31. Miao, X., et al.: SI2P: A restaurant recommendation system using preference queries over incomplete information. PVLDB **9**(13), 1509–1512 (2016)
32. Mirkin, B.G.: Core Data Analysis: Summarization, Correlation, and Visualization,Second Edition. Springer (2019)
33. Park, Y., et al.: Viz-aware sampling for very large databases. In: ICDE (2016)
34. Siddiqui, T., et al.: Effortless data exploration with zenvisage: An expressive and interactive visual analytics system. PVLDB **10**(4), 457–468 (2016)
35. Tang, B., et al.: Extracting top-k insights from multi-dimensional data. In: SIGMOD (2017)
36. Vartak, M., et al.: SEEDB: efficient data-driven visualization recommendations to support visual analytics. In: PVLDB (2015)
37. Webber, W., et al.: A similarity measure for indefinite rankings. TOIS **28**(4), 20–38 (2010)
38. Wu, C., et al.: Using association rules for completing missing data. In: HIS (2004)
39. Zhang, A., et al.: Interval estimation for aggregate queries on incomplete data. JCST **34**(6), 1203–1216 (2019)