Advancing Data Cleaning Methods for Automated Visualization Tools: A Focus on Speed, Depth, and Engagement

Rachit Chadha
rchadha33@gatech.edu

Brian Sang
bsang3@gatech.edu

ABSTRACT

In the field of data analytics, automated visualization tools like Lux are essential for extracting insights from complex data, but their effectiveness is often limited by data quality. This paper explores the intersection of data cleaning and automated visualization, particularly how various cleaning methods affect Lux's visualization outputs. We conducted a comparative analysis using datasets like the US Census and EEG data, evaluating the impact of different cleaning techniques, including basic methods like median imputation and outlier removal, and advanced techniques like AutoClean and HoloClean, on Lux's visualization recommendations. A key contribution of this study is the introduction of a DataCleaner class, an innovative tool designed to enhance the data cleaning process in conjunction with visualization systems like Lux. This class offers user-friendly cleaning options and visual feedback, aiding users in making informed data cleaning decisions. Our findings indicate that data cleaning significantly affects the visual and analytical quality of Lux's outputs, with varying impacts depending on the method used. A user study confirmed the efficiency and user-friendliness of our approach, demonstrating its potential to improve decision-making in data pre-processing. This research fills a crucial gap in the use of automated visualization tools, underscoring the importance of data cleaning in improving the accuracy and relevance of visual data representations. It sets the stage for future research aimed at maximizing the confidence of automated visualization tools with real-world data complexities.

1 INTRODUCTION

Data visualization tools play a pivotal role in contemporary data analytics by offering intuitive insights from complex datasets. One such tool, Lux, stands out for its unique capability to automatically recommend 'interesting' visualizations from pandas datasets. However, real-world datasets often come with a caveat - they are inherently messy or 'dirty'. The key question arises: How does the cleanliness of data influence the quality and relevance of visualizations suggested by Lux? While the importance of clean data is universally recognized in analytics, there's a gap in understanding its specific impact on automated visualization recommendations like those in Lux.

In the prevailing landscape of data visualization, most tools, including Lux, assume a certain level of data hygiene. The current research, such as *Lux: Always-on Visualization Recommendations for Exploratory Dataframe Workflows* [8], largely revolves around the mechanics of visualization recommendation, but there's an implicit assumption that the data is pre-processed and free from discrepancies. The challenge remains that, in practice, data scientists spend a significant amount of time cleaning data, employing a myriad of methods. Yet, there's scant research on how these different cleaning methods may influence the outcome of visualization recommendations, especially in systems like Lux.

Lux is an always-on visualization framework that allows users to utilize panda dataframes for exploratory analysis. It acts as a light-weight and quick wrapper that reduces the barrier of visualization data to enable exploration and discovery by recommending certain visualizations based on interestingness as shown in Figure 1.

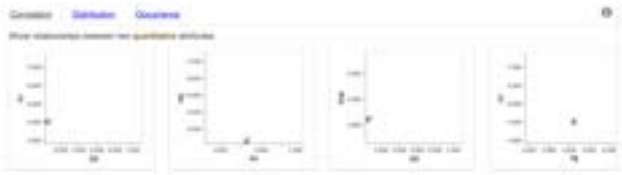

Figure 1: Example of Lux Visualization and Recommendation to Users.

Since Lux provides multiple types of charts to the users, Lux scores the interestingness of each type of plot differently, as shown in Figure 2. For bar charts, Lux ranks based on the unevenness of the chart where it looks at the chart and if it is more uneven, there's more variation in the individual bar values in the chart, which can be computed via Euclidean distance such as the L2 norm. For histograms, Lux ranks based on skewness if there's a distribution that deviates more than the normal distribution. For scatter plots, the more popular data visualization provided by Lux, Lux ranks based on monotonicity, which can be measured by the Spearman correlation coefficient along with an undisclosed significance factor.

Figure 2: Lux Interestingness Metric.

This paper delves into an exploratory invitation to address the gap of data cleaning in tandem with automatic visualization recommendation systems such as Lux by looking at the interconnected dynamics of current possible data cleaning methods and the proposed visualization recommendations from Lux. Our central hypothesis posits that various cleaning methods not only shape the visual

appeal of the representations but also impact their fundamental analytical significance. After conducting a comprehensive survey to determine the compatibility of state-of-the-art data cleaning processes with swift and user-friendly methods, the primary objective of this endeavor is to craft optimal data cleaning techniques harmoniously integrated with automated data visualization systems. Through uncovering these subtleties, our aim is to enhance decision-making in data pre-processing, ensuring that automated visualization tools realize their full potential when confronted with the inherent complexities of real-world datasets.

2 RELATED WORK

For visualizations, there are many tools that can be used for automatic data visualizations and recommendations.

2.1 Data Visualization and Recommendation

In particular, there are many data visualizations and recommendations that users utilize such as Tableau and PowerBI [13, 14], which both offer easy to use interactive interfaces for visualization construction along with the options to let users browse through via visualization recommendations. However, the data needs to be inputted at a separately and requires data preparation to be easily viewed. Data visualization and recommendation systems also extend past stand alone applications as there are now tools to be used in SQL, where the data is already in a database. A popular tool used and has inspired many current data visualization and recommendation tools is SEEDB, a visualization recommendation engine to provide fast visuals analysis by providing interesting visualizations for trends and recommends based on usefulness and interestingness [15]. However, even though SQL is a popular tool with data analysts, there has been a climb of data analysts utilizing python with panda dataframes to be able to work with visualization as it allows for data preparation and cleaning more readily to the data analysts. There have been many open source, ready to use created by the python community such as Bamboolib, Pandas-profiling and Dataprep [2, 3, 10]. However, currently the most popular one available is Lux as it allows for always on visualization with providing users quick visualizations along with smart recommendations based on interestingness. It also gives users a lot of options in terms of showing correlation, distribution and many other options. However, Lux only provides users with one part of data analysis with data visualization and recommendation.

2.2 Data Cleaning

Another aspect of data visualization is data cleaning as data analysts report spending up to 80% of time on data cleaning [4]. There has been many research to optimize data cleaning such as AlphaClean. AlphaClean is a framework that provides users with a rich library to allow them to pick how the data cleaning parameters should be utilized in a SQL aggregate query[7]. There are also many data cleaning processes within python, such as ActiveClean, which is an iterative cleaning framework that correctly retrains machine learning model to make sure the data is cleaned from a small subset of the data set[6]. HoloClean is another popular state of the art data cleaning method that utilizes statistical inference to clean the data by utilizing a probabilistic model for data cleaning[11]. BoostClean was also created as a data cleaning method that automatically selects an ensemble of error detection and repair combinations with statistical boosting[5]

However these state of the art methods are non-ideal to be used for quick data visualizations since they are lengthy and require relatively a lot of supervision in comparison to basic data cleaning methods. These methods are also not as easy to install, which would make the overall process to data clean relatively longer and can't be used as ubiquitous.

AutoClean is an alternative data cleaning tool that utilizes basic data cleaning tools with an additional setting to automatically clean the data set[1]. It performs relatively quickly and can easily be installed by users, which makes it ideal for users of automatic data visualization tools. However, it acts as a black box as it does not tell you what data cleaning methods are used to perform on the data set and why it utilizes those data cleaning methods.

We aim to address these limitations of not just AutoClean, but also other state of the art methods in this paper to create a data cleaning method that blends with automatic data visualization tools such as Lux.

3 METHOD

In order to see the effect of possible data cleaning methods on data visualization recommendation systems, we will use real world examples of datasets, such as US Census and EEG data [9, 12] that have been previously used in previous data cleaning paper, such as HoloClean and CleanML, to determine if data cleaning has an effect on automatic data visualization. From those datasets, we utilized basic data cleaning methods such as median imputation and most frequent imputation along with outlier removal to determine if these basic data cleaning methods have an effect on the visualization recommendations provided by Lux. We will then qualitatively look at the visualizations provided by Lux and also look at the ranking changes of which data visualization to first show to determine if the basic data cleaning method has an effect. Lux's ranking system of data visualization is based on interestingness score. As mentioned previously, Lux utilizes a lot of different metrics for rankings, so we wanted to see if data cleaning has a big enough effect on the interestingness of the plots provided by Lux and if there are major changes with not just the rankings, but also the visualizations.

We then wanted to see if state of the art data cleaning methods found in prior literature can be utilized in tandem with automatic data visualization systems, specifically Lux, and to determine if there is a difference with visualization and if it can be done easily in a time efficient manner. We tested AutoClean and HoloClean on a real world dataset based on EEG data. We also compared the visualization and recommendations with the basic data cleaning methods.

We also wanted to create a tool designed to enhance the data cleaning process, particularly in conjunction with visualization systems like Lux. The primary objective of this python class is to provide a user-friendly and efficient way to visualize and assess the impact of various data cleaning methods before final selection and application. This approach is twofold:

(1) **Integration of a Range of Data Cleaning Methods:** The DataCleaner class offers a suite of data cleaning options, including 'Median + Most Frequent', 'KNN Imputation', and various outlier removal techniques. This range of methods caters to different data types and scenarios, allowing users to automate the cleaning process while also having the flexibility to explore how each method impacts their specific dataset.
(2) **Visual Feedback and User Interaction:** A key feature of the DataCleaner class is its ability to provide visual feedback on the effects of different cleaning methods. This aligns with Lux's visual-first approach, aiding users in understanding the implications of each cleaning method on their data. The class is designed to incorporate user feedback and interaction, enabling users to see firsthand how different methods handle outliers, impute missing values, and ultimately influence the dataset. This interactive process allows users to make informed decisions, ensuring that the final visualizations are not only clean and accurate but also customized to the dataset's unique characteristics.

In the final development phase of the DataCleaner class, our focus is on creating an advanced data cleaning approach that seamlessly integrates with Lux, enhancing user experience in data visualization. This final version will incorporate sophisticated data cleaning techniques while still offering user customization, catering to the complexities of diverse datasets. We are committed to optimizing both efficiency and usability, ensuring the process is effective and time-efficient, particularly for large datasets. A user-centric design is at the forefront, with the class being tailored to user preferences and feedback, making data cleaning intuitive and user-friendly. Our goal with this methodology is to seamlessly bridge the gap between intricate data cleaning tasks and intuitive data visualization, thereby crafting a comprehensive tool that elevates the quality and depth of data analysis.We also intend to utilize pruning to show the most interesting plots as determined by metrics similar to Lux interestingness score.

With all of this in consideration, we created a DataCleaner class that significantly enhances the data cleaning process, particularly when used in conjunction with visualization systems like Lux. It is capable of identifying a range of data issues, such as missing values, duplicate rows, incorrect data types, and outliers. This class stands out for its automated yet customizable approach to data cleaning, providing users with a variety of methods such as 'Median + Most Frequent', 'KNN Imputation', and different techniques for outlier removal. This flexibility is key, as it allows users not only to automate the cleaning process but also to visually explore the impact of different cleaning methods on their data. By offering visual feedback on how each method alters the dataset, the class aligns seamlessly with Lux's visual-first approach, aiding users in understanding and selecting the most suitable cleaning method for their specific analysis needs.

4 EVALUATION

Our thesis examines the influence of different data cleaning approaches on the visual and analytical aspects of data representations, with the goal of creating efficient, user-friendly methods that seamlessly blend with automated data visualization systems.

4.1 Data Cleaning Effect

Throughout our work, we first wanted to see if basic data cleaning had an effect on the visualizations and visualization rankings provided by Lux on real world data sets, such as EEG data. Through basic data cleaning such as median imputation and outlier removal, we have seen a sudden shift in the visualizations along with its rankings as shown in Figure 3.

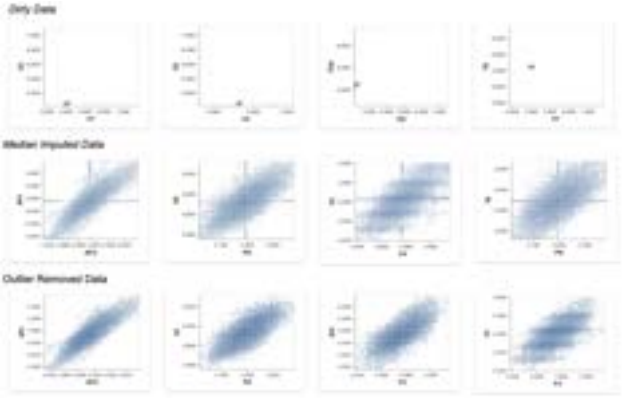

Figure 3: EEG Dataset Visualization and Ranking Shift (Raw/Dirty vs Median Imputed vs Outliers Removal)

The sudden shift from basic data cleaning methods means that not only does data cleaning have an effect, but also the type of data cleaning has an impact on the visualization changes and affects the user's perception of the graphs provided by Lux.

We also tested state of the art data cleaning methods such as AutoClean and HoloClean, where we wanted to see if these data cleaning methods can be used intuitively with Lux. After utilizing AutoClean on the EEG data along with USCensus data, we have seen a visualization and recommendation shift from not just the uncleaned data but also from the data with basic data cleaning as shown in Figure 3.

Even though AutoClean provides an automatic way to clean the real world data sets, the bias of the visualizations and recommendations still have an effect as shown in Figure 4, 5 and 6. In addition for the US Census data set, only three visualizations were recommend to the user with both types of data cleaning methods, which might not be ideal as it limits the user's options. We also see similar changes in visualizations and recommendations in the occurrence option provided by Lux as shown in Figure 6.

Therefore, it is ideal to allow the users to pick which data cleaning method they prefer. We also don't see much difference in terms of basic data cleaning when selecting the occurrence option of Lux,

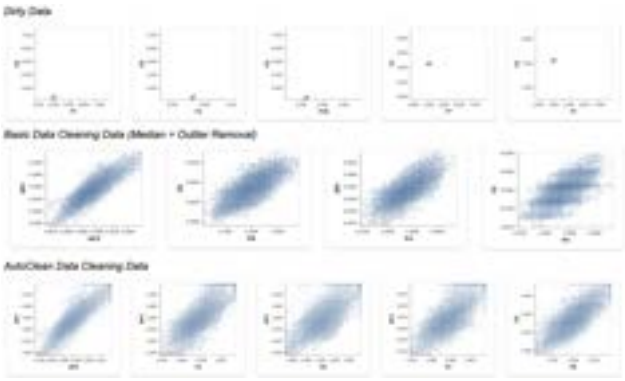

Figure 4: EEG Correlations Visualization and Ranking Shift (Raw/Dirty vs Basic Cleaning vs Automatic Cleaning)

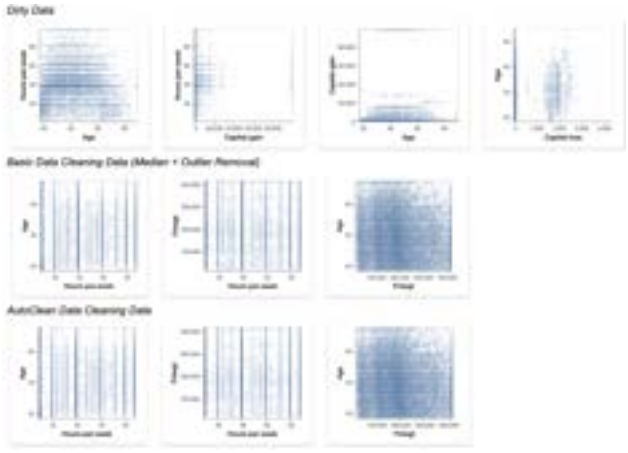

Figure 5: USCensus Correlations Visualization and Ranking Shift (Raw/Dirty vs Basic Cleaning vs Automatic Cleaning)

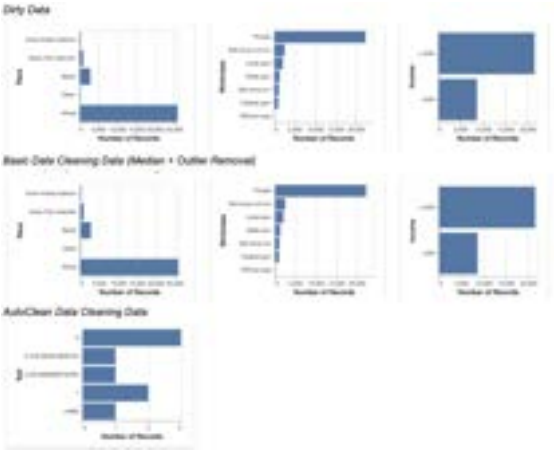

Figure 6: USCensus Occurrence Visualization and Ranking Shift (Raw/Dirty vs Basic Cleaning vs Automatic Cleaning)

but we do see a major difference when looking at Lux's recommendations after AutoClean. These major differences in bias show as to why data cleaning for Lux needs to be transparent to the user as it can often lead to confusion as to why certain data visualizations after data cleaning are lost as it provides additional insight about the dataset.

HoloClean, another state of the art data cleaning method, is an automatic data cleaning method that uses semi-supervision. Even though HoloClean has shown success for data cleaning, the intent for HoloClean to be used with quick, automatic data visualizations systems like Lux are non-ideal as HoloClean is not easy to use and has a learning curve with specific required constraint files to be run in Python 3.6 environment or prior.

We have also shown we can calculate interestingness and match with the visualizations and recommendations provided by Lux on the EEG data set. We have shown that the monotoncity score of the scatter plots of dirty and outlier removal data cleaning match with Lux ranking in Figure 7 and 8, respectively.

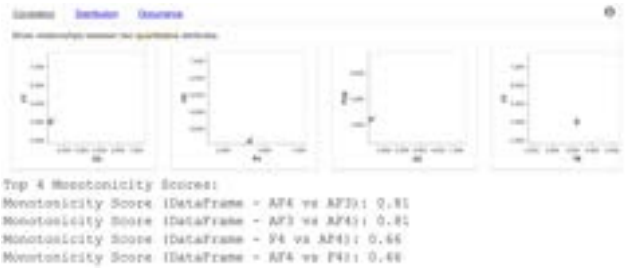

Figure 7: Monotonicty Raw Data

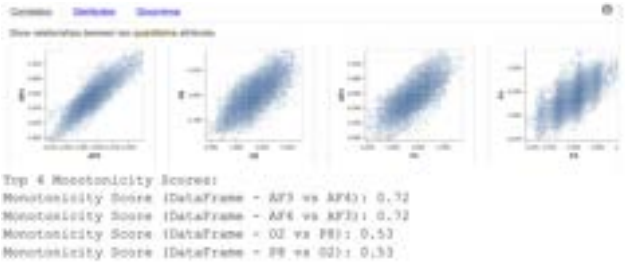

Figure 8: Monotoncity Outlier Removed

We have also shown that the unevenness score of the bar charts on dirty and outlier removal data cleaning match with Lux ranking in Figure 9 and 10.
We therefore plan to create a data cleaning method that enables users to easily select their preferred data cleaning approach, while also ensuring that only the most relevant and effective methods are presented for their specific data set. This user-centric design will enhance the overall data cleaning experience, making it more intuitive and tailored to individual needs.

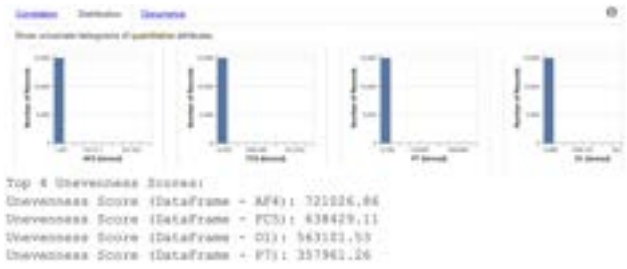

Figure 9: Unevenness Raw Data

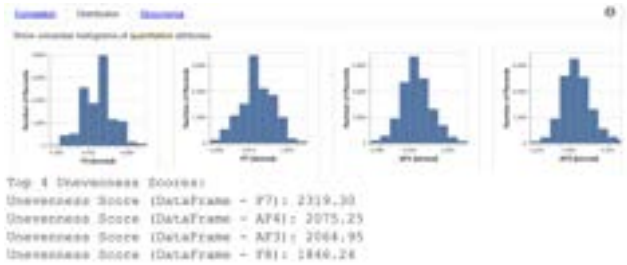

Figure 10: Unevenness Outlier Removed

4.2 Our Approach

We want to show how our data cleaning method of the DataCleaner class works in tandem with Lux and look at runtime along with general usability. Figure 11 below, shows the structure of our DataCleaner Class.

```
class DataCleaner:
    def __init__(self, df):
        # Initialize the class with a DataFrame and detect issues
        self._df = df
        self.issues = self._detect_issues()
        self._print_issues()
        self.methods = [
            "Median + Most Frequent","KNN Imputation",
            "Outlier Removal with Median Imputation",
            "Outlier Removal Ignoring Missing",
            "Outlier Removal",
        ]
    @property
    def df(self):
        # Property to get the DataFrame
        return self._df

    @df.setter
    def df(self, new_df):
        # Setter to update the DataFrame and re-detect issues
        self._df = new_df
        self.issues = self._detect_issues()
        self._print_issues()

    def _print_issues(self):
        # Print detected data issues

    def _detect_issues(self):
        # Detect various data issues like missing values, duplicates, etc.

    def apply_cleaning_method(self, df, method):
        # Apply specified data cleaning methods to the DataFrame

    def _calculate_interestingness(self, df, col1, col2):
        # Calculate the interestingness score for a pair of columns

    def visualize_effect_of_methods(self):
        # Visualize the effect of cleaning methods, pruning less interesting visualizations

    def visualize_all_methods(self):
        # Visualize the effect of all cleaning methods, showing interestingness scores
```

Figure 11: Code snippet of the DataCleaner Class structure

The design of the DataCleaner class, which emphasizes user feedback and engagement, is particularly advantageous in the realm of data cleaning—a task that often requires customization based on the unique demands of the analysis. Users are able to witness firsthand the effects of various cleaning methods, including the handling of outliers and the imputation of missing values. This level of visibility and interactivity empowers users to make more informed decisions regarding their data cleaning strategy. As a result, the final visualizations produced are not only clean and accurate but also finely tuned to the distinct characteristics of the data set as shown in Figure 12. In essence, the DataCleaner class presents a user-friendly, insightful, and adaptable tool, effectively bridging the gap between the complexities of data cleaning and the intuitiveness of data visualization. Figure 12, 13 and 14 shows the example usage and problem identification of the class, emphasizing how data cleaning choices can drastically affect visualization outcomes

```
#Load your dataset
df = pd.read_csv('/raw.csv')

#detect issues
data_cleaner = DataCleaner(df2)
```

```
Issues detected:
 - Missing values detected.
 - Outliers detected.
 - High cardinality in column Education.
 - High cardinality in column Occupation.
 - High cardinality in column Native-country.
```

Figure 12: Example Usage Step 1: Data cleaning issue identification on US Census Data

```
# Visualize based on interestingness
data_cleaner.visualize_effect_of_methods()
```

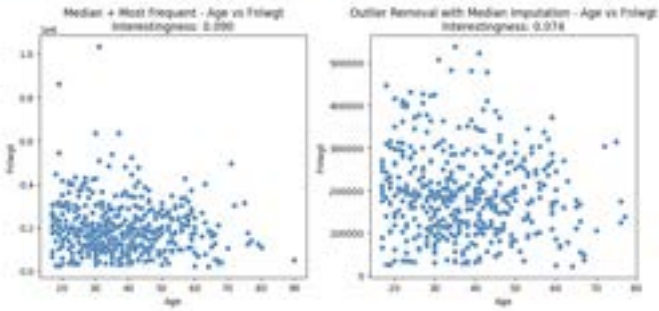

Figure 13: Example Usage Step 2: Cleaning Method Visualized and pruned based on Interestingness (US Census Dataset)

The DataCleaner class integrates an innovative feature to calculate 'interestingness' scores for different data visualizations, resonating well with the visualization recommendations typically offered by tools like Lux. This is especially evident in datasets such as US

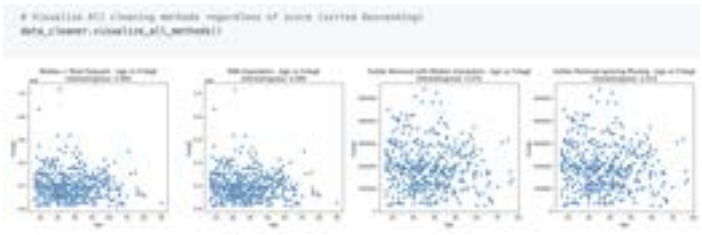

Figure 14: Example Usage Step 3: All Cleaning Method Visualized regardless of Interestingness (US Census Dataset)

Census and EEG, where understanding complex patterns is crucial. Figure 13 illustrates how the class evaluates the monotonicity score of scatter plots, effectively pruning visuals with similar interestingness levels to highlight the most significant ones. In a similar vein, Figure 14 demonstrates the use of the DataCleaner class to compare monotonicity scores across scatter plots derived from all cleaning methods, with the results sorted in descending order of interestingness.

At its core, the DataCleaner class embodies a harmonious blend of intuitiveness and analytical depth in data processing. It adeptly navigates through the intricacies of data cleaning, simultaneously enhancing the overall data visualization experience. Figures 12, 13 and 14 exemplify the class's practical application, showcasing its ability to not only identify problems but also provide insightful solutions. These examples underscore the class's role as a user-friendly, adaptable, and insightful tool in the data analysis landscape, making it an invaluable asset for both novice and expert data practitioners.

4.3 User Study

In order to determine if this proposed data cleaning approach is feasible for Lux users, we conducted a user study of the overall process, while also looking at the runtime, to determine if this data cleaning approach is efficient and user-friendly that seamlessly blends with automatic data visualization systems, specifically Lux.

We did a user study to determine how the users, which are three college students studying computer science with minimal exposure to Lux, interacted with Lux only and the same users were also asked to utilize Lux with the proposed data cleaning method. The users were asked to work with the US Census data set, which is known to be a dirty data set, and the users were asked to pick the most interesting data. The users were not informed that the US Census data set is considered dirty. We then determined if the user selected graphs are different from Lux in comparison with Lux and data cleaning. Throughout the user study, the users were then asked to talk while performing tasks to inform what they were thinking about the data cleaning approach. After the testing protocol, the users were then asked on a scale of their ease of use of both Lux and Lux with data cleaning. They were also asked on a scale how confident they are with the graphs chosen on how interesting it is. We also measured runtime on all aspects including the time it took for users to set up the data cleaning process and the amount of time that the users used to select a graph.

The results of the user study comparing Lux alone and Lux integrated with the proposed data cleaning method revealed notable improvements in users' data exploration experience. When using Lux alone, participants tended to select graphs that captured broad trends but often overlooked subtle anomalies in the dirty US Census dataset. However, when utilizing Lux with the integrated data cleaning process, users consistently demonstrated a deeper understanding of the dataset, reflected in their graph selections. The selected visualizations showed improved clarity and relevance, indicating the efficacy of the data cleaning approach in enhancing the quality of insights. While users initially rated Lux with data cleaning slightly lower in terms of ease of use due to the learning curve associated with the integrated features, they reported increased comfort and efficiency over time. The confidence ratings in selected graphs significantly rose when using Lux with data cleaning, reflecting users' enhanced trust in the validity of visualizations derived from cleaned data. Despite an initial setup time increase of 5 minutes for the data cleaning process, users' overall time spent on selecting a graph remained consistent at 5 minutes, with a reported decrease in setup time as users became more familiar with the integrated features. In conclusion, the combined Lux and data cleaning approach proved to be a powerful and user-friendly solution, significantly improving data exploration, visualization quality, and user confidence. From this user study, we can determine if the data cleaning approach we proposed is easy to use and gives users more confidence in their data.

5 DISCUSSION

5.1 Impact of Data Cleaning on User Preference

Data cleaning impacts visualizations and recommendations significantly, and this effect varies depending on the dataset. In time series data such as EEG, outliers might be crucial as they can represent significant events or anomalies. The presence of outliers can lead to notable differences in the interpretation of data. This aspect was particularly evident in our studies with EEG data, where the variance in outliers was prominent.

5.2 Adapting to User Preferences and Datasets

Users desire confidence in their data cleaning methods, especially when dealing with complex datasets like the US Census data. Our study highlighted that even in non-time series datasets, the preference for retaining or removing outliers can significantly influence the results. Providing users with options to customize their data cleaning approach, such as choosing whether to consider outliers, can enhance the utility and effectiveness of data cleaning tools.

5.3 Future Directions and User Studies

Our findings suggest a need for expanding user studies to include a broader spectrum of data analysts. By examining how various user groups interact with data cleaning tools, we can gain valuable insights into improving these tools for different levels of expertise and application contexts. Future studies should focus on how users from diverse backgrounds and with varying degrees of experience utilize data cleaning methods in their analysis workflows.

5.4 Improving User Experience and Methods

The DataCleaner class's integration with Lux offers a promising approach to data cleaning, but there is potential for further enhancing user experience. Reducing the initial learning curve and introducing more intuitive interfaces can make the tool more accessible, especially for users new to data cleaning or Lux.

6 CONCLUSION

In conclusion, our investigation reveals that conventional state-of-the-art data cleaning methods may not fully align with the needs of automatic data visualization and recommendation tools, particularly when applied to real-world datasets like the US Census and EEG data. Despite this, data cleaning undeniably influences the outcomes of such tools, notably in the case of Lux. To address this gap, we have developed the DataCleaner class, an innovative solution designed to streamline the data cleaning process in synergy with automatic data visualization systems like Lux. As evidenced by our user study, the DataCleaner class not only bolsters user confidence in utilizing Lux but also enhances the efficiency of achieving the primary objective of rapid and accurate data visualizations. These visualizations provide users with insightful glimpses into their datasets, facilitating better decision-making.

The DataCleaner class, through its user-centric approach and innovative capabilities, has emerged as a valuable tool in data exploration and visualization. Nonetheless, our findings also highlight areas for potential enhancement, particularly in refining the user experience and tailoring the tool to meet the varied needs of different users and datasets. Future developments will focus on iterative improvements based on user feedback and on conducting more comprehensive studies encompassing a broader spectrum of datasets and user demographics. This expansion is crucial for further refining the DataCleaner class and solidifying its role as an indispensable asset in data analysis, particularly for users who seek a harmonious blend of data cleaning efficiency and insightful visualization.

REFERENCES

[1] AutoClean. 2022. AutoClean.
[2] bamboolib. 2020. bamboolib.
[3] dataprep. 2023. Dataprep.
[4] Joseph M Hellerstein. 2008. Quantitative Data Cleaning for Large Databases. http://db.cs.berkeley.edu/jmh
[5] Sanjay Krishnan, Michael J. Franklin, Ken Goldberg, and Eugene Wu. 2017. BoostClean: Automated Error Detection and Repair for Machine Learning. (11 2017). http://arxiv.org/abs/1711.01299
[6] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J Franklin, and Ken Goldberg. 2015. ActiveClean: Interactive Data Cleaning For Statistical Modeling.
[7] Sanjay Krishnan and Eugene Wu. 2022. AlphaClean: Automatic Generation of Data Cleaning Pipelines. *IEEE International Conference on Program Comprehension* 2022-March, 36–47. https://doi.org/10.1145/nnnnnnn.nnnnnnn
[8] Doris Jung Lin Lee, Dixin Tang, Kunal Agarwal, Thyne Boonmark, Caitlyn Chen, Jake Kang, Ujjaini Mukhopadhyay, Jerry Song, Micah Yong, Marti A. Hearst, and Aditya G. Parameswaran. 2021. Lux: Always-on Visualization Recommendations for Exploratory Dataframe Workflows. *Proceedings of the VLDB Endowment* 15, 727–738. Issue 3. https://doi.org/10.14778/3494124.3494151
[9] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. 2019. CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks. (4 2019). http://arxiv.org/abs/1904.09483
[10] pandasprofiling. 2023. Pandas-profiling.
[11] Theodoros Rekatsinas, Xu Chu, Ihab F Ilyas, and Christopher Ré. 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference.
[12] Theodoros Rekatsinas, Xu Chu, Ihab F Ilyas, and Christopher Ré. 2150. HoloClean: Holistic Data Repairs with Probabilistic Inference.
[13] Tableau. 2023. Tableau.
[14] BI Tools. 2023. PowerBI:Interactive data visualization.
[15] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. 2015. SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics.