Improving Visualization Interpretation Using Counterfactuals
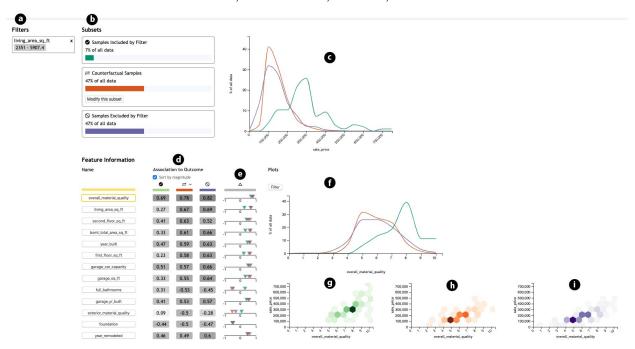
Smiti Kaul, David Borland, Nan Cao, David Gotz

Fig. 1: Counterfactual visualization tool. The user has applied a filter constraint (a) to a multidimensional dataset and is shown the resulting included, counterfactual, and excluded subsets (b), their outcome distributions (c), and other feature information (d–i).

Abstract— Complex, high-dimensional data is used in a wide range of domains to explore problems and make decisions. Analysis of high-dimensional data, however, is vulnerable to the hidden influence of confounding variables, especially as users apply ad hoc filtering operations to visualize only specific subsets of an entire dataset. Thus, visual data-driven analysis can mislead users and encourage mistaken assumptions about causality or the strength of relationships between features. This work introduces a novel visual approach designed to reveal the presence of confounding variables via counterfactual possibilities during visual data analysis. It is implemented in CoFact, an interactive visualization prototype that determines and visualizes *counterfactual subsets* to better support user exploration of feature relationships. Using publicly available datasets, we conducted a controlled user study to demonstrate the effectiveness of our approach; the results indicate that users exposed to counterfactual visualizations formed more careful judgments about feature-to-outcome relationships.

Index Terms—visualization, counterfactuals, human-computer interaction, human-centered computing, empirical study

---

1 INTRODUCTION

Supporting user inference and decision-making is one of the primary goals of information visualization, and data visualization systems therefore offer a variety of intuitive data representations and interactive tools to assist users. Visualization is used in a wide range of domains to explore problems and make decisions using large, complex, and

---

- *Smiti Kaul is with the Dept. of Computer Science at the University of North Carolina at Chapel Hill. E-mail: smiti@live.unc.edu.*
- *David Borland is with RENCI at the University of North Carolina at Chapel Hill. E-mail: borland@renci.org.*
- *Nan Cao is with the Intelligent Big Data Visualization Lab at Tongji University. E-mail: nan.cao@tongji.edu.cn*
- *David Gotz is with the School of Information and Library Science at the University of North Carolina at Chapel Hill. E-mail: gotz@unc.edu.*

high-dimensional datasets [50], with some level of trust that it will reveal important, otherwise easy-to-miss information. Visualizations do provide crucial insights, but visual data analysis can often overlook the hidden influence of confounding variables. This is especially true if users apply ad hoc filtering operations to visualize specific subsets of high-dimensional data [2]. This "overview first, zoom and filter, then details on demand" workflow [61] is found in many popular data visualization tools, such as Tableau. Although this approach is invaluable for managing large data, it may also encourage mistaken assumptions about causality or the strength of relationships between features. For example, imagine two groups of individuals, where group A is active on social media and group B is not. Presented only with visual information that shows that group A is overall unhappier, one might infer that social media activity determines individual levels of happiness. While an analyst may conduct more rigorous data analyses to reach any definite conclusions, the absence of visual cues about the effect of other attributes on the two groups can lead to unconscious assumptions about social media's relationship with happiness, which can bias subsequent analytical work. In this scenario, the theory of counterfactual thinking would urge the following questions: Could individuals be unhappy

even without social media? What are other factors that contribute to unhappiness, more or less so than social media usage does?

Fields such as causal analysis and explainable artificial intelligence have growing analytical and visual support for counterfactual thinking, but there is little to no such support for it in data exploration stages that precede causal analysis or data modeling via machine learning. We believe that confounding factors are easy to overlook during visual data exploration, as confirmed by the results presented in Section 5. Motivated to fill this gap, we developed CoFact, a novel visual approach and system prototype that can reveal the presence of confounding factors during earlier stages of visual data analysis. CoFact enables users to interactively explore data, perform filtering operations, and analyze *counterfactual subsets* to help guard against potentially erroneous conclusions about feature-to-outcome relationships. The key contributions in this paper are summarized as follows:

- **Approach:** We present a novel way of visualizing and analyzing counterfactual subsets to investigate the influence of confounding factors within large and complex datasets.

- **Prototype:** We introduce CoFact, a visualization system prototype to reveal the influence of confounding variables and improve user decision-making during earlier stages of data analysis.

- **Evaluation:** We demonstrate the effectiveness of our approach as implemented within CoFact through a controlled user study and interviews with 30 participants. Results show that the counterfactual visualizations significantly influenced user inference, reducing user confidence in weak feature-to-outcome relationships while confirming higher confidence in stronger feature-to-outcome relationships.

2 RELATED WORK

The counterfactual-based approach to visualization presented in this paper builds upon prior research in a number of closely-related areas. This section discusses work in the areas of causal analysis, explainable machine learning, and data subset creation and analysis.

2.1 Counterfactual Thinking for Causal Analysis

Early work on causality theory includes that by Pearl [51, 52], Spirtes [64, 65], and others. Causality theory rests on counterfactual thinking [25, 39]: if A causes B, then in an alternative, "counterfactual" scenario where A does not occur, B will not occur. Counterfactual thinking also asks us to investigate possible scenarios in which A does not occur but B occurs nonetheless. Byrne [4] adds that counterfactuals can amplify causal judgement, since knowing that an alternative scenario that eliminates A would not lead to B would amplify one's judgement of a causal relationship between A and B. Knowing that an alternative scenario eliminates A but also leads to B would weaken confidence in the influence of A on B. We revisit this idea in Section 3.1 to explain the value of visualizing counterfactual subsets.

The use of large-scale data for causal inference and analysis is widespread, as researchers and analysts across domains seek to understand how data attributes influence certain outcomes [70]. Many approaches exist to identify and model causal relationships in observational data [44, 53, 59], and several visualization systems support such analyses with suites of data visualization and interaction tools. Traditional visualizations include directed acyclic graph (DAG) layouts and Hasse diagrams [27]. Researchers have also proposed alternatives, such as Growing Squares [16] and Growing Polygons [15], to enhance typical DAGs. The Visual Causality Analyst offers 2D graph views and statistical parameters to reveal possible causal influences of variables [70]. Others have introduced animations illustrating causal relationships [29] and a visual causal analysis system for hypothesis generation and evaluation [5]. While determining causality is not this work's primary goal, casual analysis does provide the context within which CoFact aims to provide additional insight. In contrast to past approaches, CoFact does not depend on or depict abstract DAGs. Instead, it visualizes counterfactual subsets, as described later, to facilitate analysis during data exploration.

2.2 Counterfactual Explanations in Machine Learning

Counterfactual thinking has gained increased recent attention in explainable machine learning research. Similar to counterfactual thinking for causal analysis [39], counterfactual explanations explore what modifications in the data would lead to an alternative prediction by a machine learning model [47, 69]. Several techniques exist for generating counterfactual explanations [19, 24, 31, 41, 43, 63]. One example is DATE, which focuses on optimization and tree-based models [42]. Most techniques for counterfactual explanation generation focus on deep neural networks [40, 66], while LIME and DiCE are model-agnostic tools [49, 57].

Visualization systems for counterfactual explanations have also been developed to present actionable insights. For example, DECE supports comparison of subsets' counterfactual examples [6]. Other systems include Prospector [37] and RuleMatrix [48], both of which are model-agnostic. The What-If Tool [71] enables interactive exploration of machine learning models to help users find the nearest "counterfactual" data point. ViCE [20] visualizes the minimum modification required to change a model's prediction, and [21] introduces counterfactual visual explanations for image data. Lastly, [54] addresses shortcomings of counterfactual explanations and proposes FACE to find "feasible paths" between a subject's current and desired states. While such systems exist to help explore counterfactual possibilities in machine learning, counterfactual visualization support for data analysis in pre-modelling stages is rare.

2.3 Data Subsets and Counterfactual Possibilities

This paper's focus is on the data exploration stages that precede causal analysis and modeling, during which data filtering and subset creation are common preliminary steps to examine the influence of certain features on an outcome [35, 46]. Existing systems employ various interaction techniques to help users create and analyze data subsets [18, 33, 73], while others have proposed algorithms for automated feature selection [58, 67]. Visualization methods often employ correlation analysis and use scatter plots and heat maps to display feature information. Researchers have also proposed a range of geometrically transformed displays [8, 10, 28, 32, 34, 36, 38, 60] and radial visualization methods [12, 13] for feature evaluation. An example is [14], which adapts the RadViz technique [22, 23] to project attributes onto a 2D space. Others offer related approaches to dimension reduction for feature set evaluation [7, 9, 26].

Researchers have noted the uncertainty and bias problems inherent in making inferences based upon visual representations of user-defined subsets [1]. Additionally, visual data exploration is often undertaken by analysts with little a priori knowledge about the data features, lending them more vulnerable to whatever algorithmic or statistical biases are visualized [14]. While the tools mentioned above help users examine data features and their relationships, they lack explicit visualizations of counterfactual possibilities that might illuminate confounding factors. Thus, in this paper, we augment standard data subset creation, analysis, and visualization techniques by visualizing counterfactual subsets as described in Section 3. The goal of visualizing counterfactual data is to protect users from spurious assumptions about causal relationships until causality can be analytically established during later stages of data analysis.

3 COFACT

This section describes the proposed counterfactual approach to visual data analysis (Section 3.1) and the visual interface of CoFact (Section 3.3). Two usage scenarios that demonstrate the approach and interface are detailed in Section 3.4.

3.1 Counterfactual Approach

The primary goal of CoFact is to caution users against false inferences about a data variable's causal effect. To this end, CoFact visually alerts users to possible confounding factors during data exploration, which typically arise when they perform ad hoc filtering operations to analyze relationships between data variables. Key conceptual components of this visualization system include the *counterfactual subset* and *filter strength* as described below.

3.1.1 Counterfactual Subset

Consider the social media example introduced in Section 1. Typically, upon applying a filter constraint such as *"social media status = active"* on a dataset of individuals, two subsets are created: the included subset of individuals who match the filter criterion (people active on social media) and the excluded subset of individuals who do not match it (people inactive on social media). We propose a third, counterfactual subset comprising data samples that do not match the filter constraint but are similar to the included subset in other ways. Filtering thus creates three distinct subsets: (1) **the included subset (IN)**, (2) **the counterfactual subset (CF)**, which does not match the filter constraint but is similar to the included subset across other features, and (3) **the excluded subset (EX)** which does not match the filter constraint and also differs substantially from the included subset across other features.

In the social media example, subset CF would include *inactive* individuals who are *similar* to active individuals across other features such as age, time spent online, and overall well-being, whereas subset EX would include *inactive* individuals who are also *different* from active individuals across other features. We determine CF and EX by computing the similarity, as measured by the Euclidean distance [11], between samples $a_0 \ldots a_{n-1}$ in IN and samples $b_0 \ldots b_{m-1}$ that do not match the filter constraint. For categorical features, the distance between any two values is 1 if they are equal and 0 if they are not. For each b_j, we calculate the similarity to IN as the mean of the normalized distances to each a_i. As a default, which was used in our user study, we form CF using the top 50% most similar samples, and EX using the bottom 50%. Section 6.2 discusses other potential methods (e.g., alternative distance and similarity metrics, different subset sizes) for determining the counterfactual subset.

3.1.2 Characterizing Filters

Suppose that the data analysis goal is to understand to what extent various data features contribute to an outcome of interest. To do so, one can apply different filter constraints and observe their impact on the outcome distributions of the IN, CF, and EX subsets — o_{IN}, o_{CF}, and o_{EX} respectively. This impact, evaluated based on the extent to which the filter constraint (f) differentiates o_{IN}, o_{CF}, and o_{EX}, exists on a spectrum. At one extreme, after applying a filter constraint, o_{IN} might differ significantly from both o_{CF} and o_{EX}, even though IN and CF are similar across features other than f, whereas CF and EX are overall more dissimilar but share f. Such an effect suggests a notable influence of f on the outcome. At the other extreme, after applying a filter constraint, o_{IN} and o_{CF} might be very similar despite the difference in f, while o_{EX} differs significantly from both. Such an effect suggests a negligible influence of f on the outcome because o_{CF} and o_{EX} are different despite sharing f. Between these two extremes, a filter constraint's impact might lead to less definitive insights.

Given this spectrum, we characterize filter constraints on a scale from 0 to 1, where 0 corresponds to no impact by f (i.e., $o_{IN} = o_{CF}$) and 1 corresponds to a large impact (i.e., a maximal difference between o_{IN} and o_{CF}). To measure this difference and characterize filters accordingly, the Hellinger distance [62] is used for categorical outcomes and the Kolmogorov-Smirnov test [45] for numerical outcomes. To evaluate the proposed counterfactual approach within CoFact, we defined three categories of *Filter Strength* — weak $\leq 0.40 <$ **moderate** < 0.60 $\leq$ **strong** — to characterize a filter constraint's impact using these measures (Table 1). These thresholds are based on a commonly used categorization of effect size on a spectrum from weak to strong [17]. We conducted an empirical analysis to check that these thresholds are suitable for both the Hellinger and Kolmogorov-Smirnov test measures.

Filter Strength	Weak	Moderate	Strong
IN/CF Difference d	$0 \leq d \leq 0.40$	$0.40 < d < 0.60$	$0.60 \leq d \leq 1$

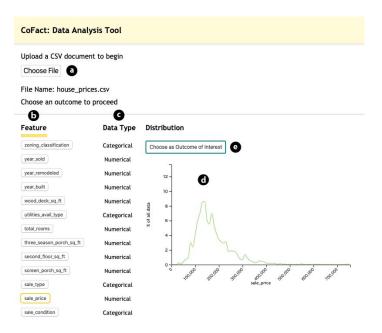Table 1: Filter strength thresholds based on IN/CF difference.

Fig. 2: CoFact landing page. The user must first choose a file (a) and then select an outcome of interest (b, e) to proceed.

Consider again the social media and happiness example. Suppose that, upon applying the filter "*social media status = active*" (f), one finds that active individuals (IN) and similar inactive individuals (CF) both have high levels of unhappiness ($o_{IN} \sim o_{CF}$), whereas inactive individuals dissimilar to active individuals (EX) have low levels of happiness ($\{o_{IN}, o_{CF}\} \nsim o_{EX}$). This filter constraint would be considered **weak**, for there is little difference between IN and CF ($d \leq 0.40$). On the other hand, suppose that, after applying f, one finds that active individuals (IN) have high levels of unhappiness ($o_{IN} \nsim \{o_{CF}, o_{EX}\}$), whereas inactive individuals both similar (CF) and dissimilar (EX) to active individuals have low levels of happiness ($o_{CF} \sim o_{EX}$). In this case the filter constraint would be considered **strong** ($d \geq 0.60$).

Among the filters used in our evaluation of CoFact that are listed in Table 2, F_{VRS} is weak because its corresponding IN/CF difference measure is 0.33. In contrast, F_S is strong because its corresponding measure is 0.69. F_{UNEMP1} is moderate, with a corresponding measure of 0.47. If a filter involves applying more than one feature constraint (e.g., "*social media status = active*" & "*age < 20*"), the IN/CF difference is calculated in the same way using the final IN, CF, and EX distributions after the compounded filters are applied, as with user study tasks $T_{Moderate1}$, T_{Weak2}, and $T_{Strong3}$ in Table 3.

3.2 Design Requirements

Based on the motivation and purpose of CoFact, three key design requirements were identified. While not derived through a formal empirical study, these are informed by the authors' long history of developing visual analytics tools in collaboration with experienced high-dimensional data users and analysts.

R1 Use feature information to choose, refine, and apply data filter constraints. Feature statistics and visualizations should help users evaluate and choose filter constraints.

R2 Understand the included, counterfactual, and excluded subsets, and, relatedly, evaluate feature-to-outcome relationships. Users should be able to understand the three subsets that result from filtering operations. Information presented about the subsets should inform user inference about the influence of a filter constraint feature on the outcome of interest.

R3 Compare feature differences across subsets. The presentation of the differences in features across subsets should further user understanding of the features and of their relationships with the outcome of interest.

3.3 Visual Interface

We designed CoFact to improve user decision-making via the proposed counterfactual approach and to support design requirements **R1–R3**. CoFact is built as an Electron application using D3 and JavaScript, and it accepts as input CSV data tables in tidy format [72].

3.3.1 Initial Analysis Configuration

The user begins by selecting a data file to load into CoFact. After loading, a summary of feature names and data types is displayed (Figure 2 (b–c)). A feature can be categorical (binary or non-dichotomous) or numerical. Clicking on any feature name button displays a corresponding distribution plot (d). Distributions are shown using line charts for continuous variables and bar charts for categorical variables. The user then selects a feature to serve as the primary outcome of interest (e). This outcome feature is used to compare subsets during the subsequent analysis.

After selecting an outcome of interest, the user is shown the first analysis page (Figure 3). The primary analytical task is to explore how the outcome feature is influenced by other features. We thus provide information about feature-to-outcome associations. The association measures described next are intended to guide users, however they are not key components of the proposed counterfactual approach. Provided both the association and counterfactual information, users may engage in an iterative process in which they compare different sets of information and modify their analyses accordingly. The visualizations in CoFact present one specific implementation, however the general counterfactual approach is not limited to it.

A list of feature buttons is displayed in the "Name" column (a), which users can sort alphabetically in ascending or descending order. The "Association to Outcome" column (b) shows how strongly a feature is associated with the outcome on a $[-1, 1]$ or $[0, 1]$ scale depending on the feature types. If the feature and outcome are both numerical, this value is the Pearson correlation coefficient. If the feature or outcome is categorical and the other is numerical, multiple regression R^2 is used. If both the feature and outcome are categorical, Cramér's V is used to determine the association-to-outcome value. In our user study (Section 4), participants viewed only correlation measures for simplicity. Plots to the right of each association value graph it on a number line (c). Users can click the green bar above the association-to-outcome column to sort in ascending or descending order. The "Sort by magnitude" checkbox enables users to ignore the signs of any correlation values.

The next analytical step is to choose a filter constraint to narrow the focus of the analysis. When a feature button (a) is clicked, two plots display detailed information about the selected feature. The first plot (d) shows the feature's distribution. Users can click and drag to select a range of feature values to use as a filter constraint. A line chart is used for numerical features and a bar chart for categorical features. The second plot shows the selected feature's relationship to the outcome of interest (Figure 3 (e)). When the selected feature is categorical and the outcome is numerical, we use a violin plot. When both the selected feature and the outcome are categorical, we use a heat map. When both the selected feature and outcome are numerical, a hex map is used as shown in Figure 3 (e).

These visualizations help users decide upon a filter (**R1**). After selecting a feature and choosing the desired value via the distribution plot, users can click the "Filter" button (Figure 3 (f)) to perform a filtering operation.

3.3.2 Subset Visualization

Applying a filter leads to the main analysis page displayed in Figure 1. The filter constraint is listed in the "Filters" column (a). The "Subsets" section (b) on the right shows the three subsets created by filtering: the included (IN) ●, counterfactual (CF) ●, and excluded (EX) ● subsets. The colored bars show what percentage of all data makes up the given subset. Plot (c) shows the outcome distributions of each of the three subsets. For example, the orange line shows what percentage of CF has certain outcome values. These visualizations support **R2** by enabling

users to compare outcome distributions across subsets to evaluate the applied filter's strength with respect to the outcome.

3.3.3 Comparative Feature Information

In Figure 1 (d), "Association to Outcome" columns now display association values for IN, CF, and EX separately. The user can sort each column in ascending or descending order. Column (e) graphs the association values for each subset on a number line. Plot (f) graphs the selected feature's distributions for each of the three subsets. Below this graph, there are three separate feature vs. outcome plots for IN (g), CF (h), and EX (i). Taken together, the feature information and visualizations provided for each individual subset help users further explore features, edit existing filter constraints, and add new ones to the list of filters (a). These components support **R3**.

3.4 Usage Scenarios

This section describes two usage scenarios to illustrate the concepts from Section 3.1 using the visual interface introduced in Section 3.3. User study participants engaged with these and six other scenarios as described in Section 4. These two scenarios use a dataset which includes information related to criminal recidivism [55]. The outcome of interest is *Recidivism Within Two Years* (*two_year_recid*), a categorical feature which the user selects on the landing page (Figure 2). A value of 0 for *two_year_recid* indicates that the defendant was not arrested again for an offense within two years since the last arrest. A value of 1 indicates that they were arrested again within two years. Using the information and visualizations on the first analysis page (Figure 3), the user can now choose and apply a filter constraint.

3.4.1 Weak Filter Scenario

The user first analyzes the influence of *Violent Recidivism Risk* (*v_decile_score*) on *two_year_recid*, the outcome of interest. To narrow their focus, they apply a filter constraint that *v_decile_score* should be *between 6 and 10*, which are relatively high risk scores. This operation creates distinct subsets to be visualized, similar to Figure 1 (which uses a numerical outcome). Figure 4 displays the outcome distributions for each subset. Figure 4 (a) only visualizes defendants with high risk scores ● and those with low risk scores ●; the counterfactual subset is empty. These visualized subsets are denoted IN and EX_{control}, respectively. Figure 4 (b), on the other hand, includes a non-empty counterfactual subset (CF ●), and EX ● is distinct from CF. IN ● is identical in (a) and (b). Looking only at plot (a), the user is likely to have high confidence in *v_decile_score*'s influence on *two_year_recid*, as there is a significant visual difference in the outcome distributions of the two subsets.

Figure 4 (b) offers contradicting information. Here, CF comprises defendants with low scores who are similar to IN in other ways. Individuals in EX still have low recidivism scores but are also more different from IN across other features than are individuals in CF. This plot shows that CF, despite having low scores, has a similar outcome distribution as IN (defendants with high scores) and a visually distinct outcome distribution from EX. This visualization alerts the user to possible confounding features that influence the outcome more strongly than *v_decile_score*, leading them to lower their confidence in the influence of *v_decile_score* on *two_year_recid*.

For visual reference, Figure 6 (a) shows what a weak filter scenario may look like for a different numerical outcome of interest. As with Figure 4 (b), IN ● and CF ● have similar outcome distributions, and both are different from the outcome distribution of EX ●.

3.4.2 Strong Filter Scenario

Next, the user analyzes the influence of *sex* on *two_year_recid*. To narrow their focus, they apply a filter constraint that *sex* should be *female*. This operation again creates distinct subsets for visualization. Figure 5 displays the outcome distributions for each subset. Figure 5 (a) only visualizes female (IN) ● and male (EX_{control}) ● defendants. Looking only at this plot, the user is likely to again have relatively high confidence that *sex* influences *two_year_recid*, albeit somewhat lower than
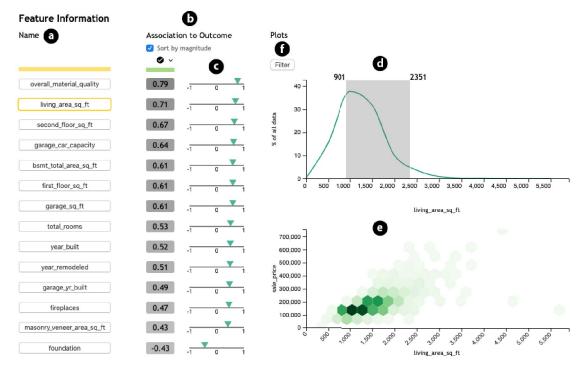
Fig. 3: First analysis page. Feature information is displayed after the user chooses an outcome of interest to facilitate filter selection.
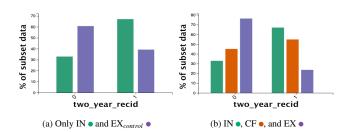

Fig. 4: Subset distributions for the outcome (*two_year_recid*) with a weak filter ($6 < v\_decile\_score < 10$).

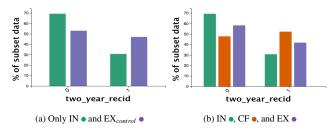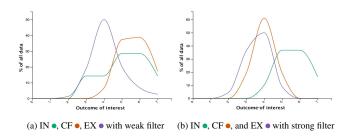Fig. 5: Subset distributions for the outcome (*two_year_recid*) with a strong filter (*sex = female*).

Fig. 6: Example subset distributions for a numerical outcome.

they did for *v_decile_score* due to a less stark difference in the subsets' outcome distributions.

In this scenario, Figure 5 (b) shows CF ●, comprising male defendants who are similar to female defendants (IN ●) in other ways. EX ● are male defendants who are more dissimilar to IN. This plot shows that CF, despite being similar to IN across other features, has a noticeably different outcome distribution than IN. CF's outcome distribution is actually close to that of EX. Thus, despite other similarities, the difference in sex differentiates outcome distributions for the female (IN) and male (CF, EX) subsets. This confirms or increases the user's high confidence in the influence of *sex* on *two_year_recid*.

Figure 6 (b) shows an example strong filter scenario for a numerical outcome. Similar to Figure 5 (b), IN ● has a distinctly different outcome distribution than both CF ● and EX ●, which are similar to each other.

4 USER STUDY

This section describes a controlled user study (n = 30) that evaluates CoFact's ability to convey counterfactual possibilities. Results from the experimental tasks and feedback from post-study interviews (Section 5) suggest that the counterfactual visualizations improved user inference about feature-to-outcome relationships without hampering system usability.

4.1 Hypotheses

The user study was designed to test the following hypotheses:

Hypothesis 1. Users exposed to the counterfactual visualizations will have lower confidence in the influence of a weak filter on the outcome of interest than users who do not view the counterfactual subset. Relatedly, users exposed to the counterfactual visualizations will have a similarly high confidence in the influence of a strong filter on the outcome of interest as users who do not view the counterfactual subset.

Hypothesis 2. Upon being exposed to the counterfactual subset, users who initially did not view the counterfactual visualizations will have decreased confidence in a weak filter's influence but will maintain relatively high confidence in a strong filter's influence.

Hypothesis 3. The added counterfactual subset visualization capability will not significantly decrease the system's usability, efficiency, or the overall quality of user experience.

Filter	Strength (IN/CF Difference)	Description	Constraint
Dataset 1: Housing Prices, **Outcome of Interest:** Sale Price			
F_{LA}	Strong (0.60)	Living area square footage	Within range 2200–5100
Dataset 2: COVID-19 State Policies, **Outcome of Interest:** Percentage of State Population with Positive Cases			
F_{SIP}	Weak (0.18)	Date state started shelter in place	On a day in April
F_{UNEMP1}	Moderate (0.47)	Percentage of state population unemployed	Within range 1–3.5
F_{RNEB}	Weak (0.25)	Date state reopened non-essential businesses	On a day in March or April
F_{WUIM}	Strong (0.68)	Weekly unemployment insurance maximum provided	Within range 240–350
F_{UNEMP2}	Weak (0.38)	Percentage of state population unemployed	Equal to or above 5
Dataset 3: Recidivism, **Outcome of Interest:** Recidivism within Two Years			
F_{VRS}	Weak (0.33)	Violent recidivism risk	Within range 6–10
F_S	Strong (0.69)	Sex	Female

Table 2: Filter constraints for the experimental data analysis tasks listed in Table 3.

	Dataset 1	Dataset 2					Dataset 3	
Task	$T_{Strong1}$	T_{Weak1}	$T_{Moderate1}$	T_{Weak2}	$T_{Strong2}$	$T_{Strong3}$	T_{Weak3}	$T_{Strong4}$
Task Strength (IN/CF Difference)	Strong (0.60)	Weak (0.18)	Moderate (0.43)	Weak (0.20)	Strong (0.68)	Strong (0.82)	Weak (0.33)	Strong (0.69)
Filter(s) Applied	F_{LA}	F_{SIP}	$F_{SIP} + F_{UNEMP1}$	$F_{SIP} + F_{RNEB}$	F_{WUIM}	$F_{WUIM} + F_{UNEMP2}$	F_{VRS}	F_S

Table 3: User study experimental data analysis tasks, using the filter constraints listed in Table 2.

Users should have decreased confidence in a weak filter after viewing the counterfactual subset because this subset reveals the presence of counterfactual explanations for differences in outcome distributions that lie beyond the difference due to the filter constraint. Users should maintain relatively high confidence in a strong filter because the resulting counterfactual subset's distribution would be distinct from that of the included subset, suggesting that the filter variable is valuable in explaining differences in outcome.

4.2 Data

We used three publicly available datasets for the user study. Dataset 1 was obtained from Kaggle and includes information (163 features, n = 1500) about houses and their sale prices [30]. Dataset 2 contains information (42 features, n = 50) related to the COVID-19 pandemic. It was formed using two separate publicly available datasets: (1) information about COVID-19 cases and deaths in U.S. states [68] and (2) U.S. state policies related to the pandemic [56]. Dataset 3, published by *ProPublica*, contains information (20 features, n = 1500) related to criminal recidivism [55].

4.3 Design and Procedure

4.3.1 Participants and Groups

We recruited 30 participants (male = 14, female = 16) via a campus-wide email, department mailing lists, and recruitment efforts within our professional network. All participants were at least 18 years old and were either pursuing or had attained a university degree. Participants belonged to a diverse range of academic and professional sectors. Under a between-subjects design, 15 were randomly assigned to the control group (C) and primarily saw visualizations for only two subsets, while 15 were assigned to the counterfactual group (CF) and viewed the counterfactual subset as well.

4.3.2 Procedures and Study Tasks

User study sessions were conducted remotely using Zoom video conferencing, and each lasted for roughly one hour. First, participants answered a pre-study questionnaire that asked them, among other questions, to rate their level of expertise on a scale from 1 (novice) to 7

(expert). Groups C and CF reported no significant difference in expertise for both general data analysis ($p = 0.50$) and visual data analysis ($p = 0.37$).

Next, we reviewed essential terms (e.g., counterfactual) and gave participants a tour of the visual interface. Participants were then provided remote control of the moderator's screen and guided through some practice data analysis tasks in CoFact, before they completed the main experimental tasks listed in Table 3. Each task's strength and corresponding IN/CF Difference measure are provided in the second column according to the methodology described in Section 3.1.2. Each participant completed 7 analysis tasks, while 23 participants (14 in group C, 9 in group CF) also had time to complete an additional task (T_{Weak2}). Each task involved applying the corresponding filter constraints detailed in Table 2 and then responding to questions **Q1–Q3** listed in Table 4. The questions asked users to describe what they observed and what inferences they drew using the visualizations.

After the pre-study questionnaire, group CF participants completed the experimental tasks. Group C participants did the same, but they also repeated T_{Weak3} and $T_{Strong4}$ after being exposed to the counterfactual subset. Despite the time constraint, this variation for group C enabled us to gather within-subjects results for at least two tasks. Section 3.4 describes the expected behaviors for T_{Weak3} and $T_{Strong4}$ in more detail. Finally, participants provided post-study feedback about the tool's usefulness and their overall experience via a questionnaire.

4.3.3 Measures

After performing each of the tasks listed in Table 3, participants reported their confidence in each filter constraint's influence on the outcome of interest on a scale from 1 (no confidence) to 7 (high confidence) in response to **Q3** in Table 4. For the post-study questionnaire, participants provided their level of agreement with eight statements related to the criteria listed in Table 5, on a scale from 1 (strongly disagree) to 7 (strongly agree).

5 RESULTS

This section reports results and feedback from the user study (Section 4). Overall, results support **Hypotheses 1–3** (Section 4.1), requirements **R1–R3** (Section 3.2), and user experience metrics **M1–M3** (Table 5).

	Counterfactual Group (CF)	Control Group (C)
[Q1] Difference in outcome distributions	**Q1-CF1:** On a scale of 1 to 7, how different are the outcome distributions for the counterfactual and included subsets? **Q1-CF2:** On a scale of 1 to 7, how different are the outcome distributions for the counterfactual and excluded subsets?	**Q1-C:** On a scale of 1 to 7, how different are the outcome distributions for the included and excluded subsets?
[Q2] Comments	What does this make you think about the filter variable's influence on the outcome?	
[Q3] Confidence	On a scale of 1 to 7, how confident are you in the filter variable's influence on the outcome?	

Table 4: Experimental questions participants responded to after each data analysis task.

Metric		Criterion
[M1] Usability	**M1.1**	Users should find the system overall easy to use.
	M1.2	It should be easy to add and remove filters.
	M1.3	It should be easy to understand the visual subset divisions that result from a filter operation.
[M2] Informativeness	**M2.1**	Users should find the system informative overall and use it to explore data in a way they would not be able to otherwise.
	M2.2	Users should find the feature-to-outcome correlations and feature distributions informative.
	M2.3	Users should find the differences in feature-to-outcome correlations and feature distributions across subsets informative.
	M2.4	Users should find the counterfactual visualization capability informative and helpful for better exploring feature-to-outcome relationships.
[M3] Efficiency		Users should analyze and understand data more efficiently using CoFact.

Table 5: Criteria to evaluate user experience.

5.1 Confidence

We evaluated self-reported confidence with a 2 (treatment group: C ● vs. CF ●) × 2 (strength: weak vs. strong) repeated measures ANOVA with treatment group as a between-subjects factor and strength as a within-subjects factor using the afex package in R. See Table 6. Post-hoc analysis was performed using estimated-marginal means with Tukey method adjustments for repeated tests. The $T_{Moderate1}$ task was removed from this analysis due to there being only a single moderate task in the experiment.

Below we analyze differences in self-reported confidence between the C and CF treatment groups with respect to overall task strength and individual tasks. We report significant results with effect sizes and confidence intervals. In general the effect sizes are moderate to strong, which provides support for **Hypotheses 1 and 2** related to participant confidence.

	F	η^2	p
TREATMENT	3.44	0.07	0.07
STRENGTH	39.33	0.35	0.00
TREATMENT:STRENGTH	6.42	0.08	0.02

Table 6: Results from the 2 (treatment group: C vs. CF) × 2 (strength: weak, strong) repeated measures ANOVA evaluating user confidence.

A significant main effect of strength was found ($F(1, 28) = 39.33$, $p < 0.001$, $\eta^2 = 0.35$). Participants were significantly more confident with the strong tasks ($M = 5.15, SD = 1.43$) compared to the weak tasks ($M = 3.77, SD = 1.67$). This main effect is quantified by the higher order significant treatment × strength interaction ($F(1, 28) = 6.42$, $p = 0.02$, $\eta^2 = 0.08$).

The CF treatment group was significantly less confident when evaluating weak tasks ($M = 3.26$, $CI = [2.75, 3.76]$) compared to all other conditions. The CF treatment group evaluating weak tasks was less confident than the CF treatment group evaluating strong tasks ($M = 5.17$,
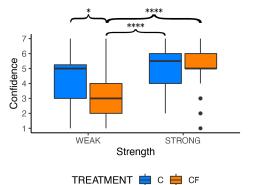

Fig. 7: Box plots of confidence for strong and weak tasks by treatment group.

$CI = [4.67, 5.67]$, $t(28) = 6.23$, $p < 0.0001$), the C treatment group evaluating weak tasks ($M = 4.32$, $CI = [3.82, 4.82]$, $t(52.8) = 3.02$, $p = 0.02$), and the C treatment group evaluating strong tasks ($M = 4.63$, $CI = [4.63, 5.63]$, $t(52.8) = 5.32$, $p < 0.0001$). No other significant results were found. See Figure 7. These results support **Hypothesis 1**, that participants exposed to the counterfactual visualizations (CF) will have lower confidence in the influence of a weak filter on the outcome of interest than participants who do not view the counterfactual subset (C), but will have a similarly high confidence in the influence of a strong filter.

	F	η^2	p
TREATMENT	1.33	0.02	0.26
TASK	8.75	0.24	0.00
TREATMENT:TASK	3.40	0.11	0.01

Table 7: Results from the 2 (treatment group: C vs. CF) × 8 (task) repeated measures ANOVA evaluating user confidence for each task.

We further investigated if there were differences between treatment groups for each of the tasks with a 2 (treatment group: C vs. CF) $\times$ 8 (task) repeated measures ANOVA (Table 7). Post-hoc analysis was again performed using estimated-marginal means with Tukey method adjustments for repeated tests. A significant main effect of task was found ($F(4.40, 92.43) = 8.75$, $p < 0.001$, $\eta^2 = 0.24$). This main inter-action was quantified by the higher-order significant treatment $\times$ task interaction ($F(4.40, 92.43) = 3.40$, $p = 0.01$, $\eta^2 = 0.11$). Post-hoc analysis was performed pair-wise comparing treatments for each task using estimated-marginal means with Tukey method adjustments for repeated tests. Significant differences were found between the C and CF treatments for T_{Weak2} ($t(144) = 2.59$, $p = 0.01$, $\eta^2 = 0.04$) and T_{Weak3} ($t(144) = 3.76$, $p = 0.0002$, $\eta^2 = 0.09$). Participants in group CF had significantly lower confidence compared to those in group C. See Figure 8.

These results further support **Hypothesis 1**. The lone exception is for task T_{Weak1}, where participants in group CF did not have significantly lower confidence than participants in group C. A possible explanation for this phenomenon relates to existing knowledge of the subject matter. As information about the COVID-19 pandemic was widely spread at the time of the user study, participants may have had difficulty separating prior knowledge from the graphical presentation of the subset outcome distributions. The plotted lines may not have fit with their understanding of the real-world role of the filter variable (F_{SIP} in Table 2) and thus confused their thinking. Participants often indicated that they did not know what to make of this scenario, and confidence was relatively low for both the C and CF groups (median value of 3).

As noted earlier, group C participants repeated a single weak task (T_{Weak3}) and strong task ($T_{Strong4}$) a second time with the counterfactual visualizations. We investigated the effect on confidence for these tasks and participants. Analysis was performed with a 2 (strength: weak vs. strong) $\times$ 2 (treatment group: C vs. CF) repeated-measures ANOVA with both strength and treatment as within-participant variables.

Significant main effects of strength ($F(1, 14) = 8.73$, $p = 0.01$, $\eta^2 = 0.08$) and treatment ($F(1, 14) = 8.20$, $p = 0.01$, $\eta^2 = 0.08$) were found. These main effects were quantified by the higher-order strength $\times$ treatment interaction ($F(1, 14) = 24.80$, $p < 0.001$, $\eta^2 = 0.21$). Post-hoc analysis was performed comparing treatment for each strength. For T_{Weak3}, participants in the CF condition had significantly lower confidence ($M = 3.13$, $SE = 0.36$, $CI = [2.4, 3.87]$) compared to those in the C condition ($M = 5.33$, $SE = 0.36$, $CI = [4.6, 6.1]$, $t(28) = 5.55$, $p < 0.0001$). No significant difference was found between the C and CF conditions in $T_{Strong4}$ ($t(28) = -1.51$, $p = 0.14$). See Figure 9. These results support **Hypothesis 2**, that user confidence would decrease for a weak filter but remain relatively similar and high for a strong filter upon viewing the counterfactual visualization.

Figure 10 displays box plots of the ratings participants provided for the differences in outcome distributions ● between the included (IN) and counterfactual (CF) subsets. It shows the corresponding confidence ● participants reported in the filters' influence on the outcome. The 4th and 10th pair of bars (T_{Weak3} and $T_{Strong4}$, respectively) display group C's responses after viewing the counterfactual visualizations. Generally, if participants reported a lower IN/CF difference rating, their confidence was lower as well. The exception is T_{Weak2}, for which participants hesitated to report high confidence without knowing how the two applied filters ($F_{SIP} + F_{RNEB}$) interacted, despite reportedly observing a notable difference in the subsets' outcome distributions.

5.2 User Experience

After completing the data analysis tasks, participants provided feedback about their experience using CoFact. Each participant rated the extent to which they agreed with the statements related to usability (**M1**), informativeness (**M2**), and efficiency (**M3**) detailed in Table 5. Figure 11 displays these responses by groups C ● and CF ●. Overall, ratings were favorable, with median ratings between 5 and 7, where 5 is "somewhat agree" that a metric quality (e.g., usability of filtering) was met and 7 is "strongly agree." Across all questions, no significant differences in rating were found between groups C and CF, supporting **Hypothesis 3**. In addition to these numeric ratings, the semi-structured

post-study interviews captured a wide range of free-form feedback. Below we summarize findings from a qualitative thematic analysis [3] of the interview responses.

Overall workflow. Participants mostly conveyed positive feedback about the system's overall usability and presentation. 22 of 30 partici-pants (13 in group C, 9 in group CF) specifically noted that the interface was straightforward and easy to use ("gave immediate visual feedback," "very intuitive and easy to use," "user-friendly") (**R2**). 12 participants noted that they appreciated the ease of adding and removing filters (**R1**). They also found useful the correlation information ("that part would be really useful," "ordering by correlation – I love that") (**R1, R3**). 28 participants answered "Yes" when asked whether or not the system helped them complete the data analysis tasks, and comments included "Yes absolutely strongly" (**R2**). One participant in group CF who indicated a "No" explained that there was a "lot of confusion."

Counterfactual visualization and interpretative assistance. Overall, participants found the counterfactual visualization valuable (**R2**). Comments included "prevents you from jumping to conclusions," "interesting way to look at data that I haven't seen," and "game changer feature." The visualization also "introduced new questions in a good way," "gives more nuance," and seemed like "the voice of reason." Group CF participants did express that more effort was required to understand the counterfactual subset ("I found myself getting a bit confused," "I had a little bit of trouble"). Several (11 of 30) asked for additional annotations and interpretation-related indicators. One said it would help if the tool could "automatically give some hints" about what the subsets mean. Another asked for "confidence intervals, particularly for the counterfactual bar graphs." Lastly, participants expressed con-cern and curiosity about how the counterfactual subset is determined, as well as that they would like to choose what features are used to determine subset similarity: "more transparency on what's considered similar features," "flexibility on how to calculate that." Some suggested that the counterfactual subset could be an optional visualization that users can turn on and off.

Customization, additional features, and potential applications. Three participants from different domain backgrounds noted that they were more familiar with other types of graphs and suggested the ability to choose how data is plotted, i.e., using a pie chart instead of a his-togram. Five mentioned the ability to customize colors. Also, while most appreciated the ability to choose filter ranges by clicking and dragging on the distribution plots, 11 wanted more precise control and the ability to type values. Two participants noted that CoFact would be a valuable teaching tool to demonstrate certain data analysis and statistical concepts. One said, "something like this would have a huge impact on an intro [computer science] course," "shows...how you can make more meaning from huge amounts of data."

6 DISCUSSION

This section discusses both key implications and limitations of the user study and CoFact (Section 6.1). Potential future directions for this work are also presented (Section 6.2).

6.1 User Study Results and Limitations

The user study sought to evaluate the three hypotheses described in Section 4.1, as well as to gather general feedback about user experience with CoFact. Overall, results (Section 5) support **Hypotheses 1–3**: the counterfactual subset visualizations improved user judgment with-out hampering usability and the overall quality of user experience. Feedback from participant interviews also gave us a more nuanced un-derstanding of this work. Generally, the visualizations helped support the primary requirements **R1–R3** (Section 3.2): participants indicated that feature information and visualizations helped them choose, apply, and refine filters, after which they were able to, in general, understand the visual subset representations and form judgments about feature-to-outcome relationships. In the future we would allow more time for unguided exploration of the user interface, specifically to more rigorously test **R1** and **R3**.

Two main limitations of the user study were (a) its time constraint and (b) the ambiguity of confidence evaluation. First, users had a

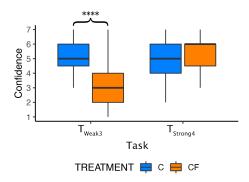Fig. 8: Box plots of confidence by treatment group for each task.

Fig. 9: Box plots of confidence for the T_{Weak3} and $T_{Strong4}$ tasks for participants who started in the C condition and then repeated these tasks in the CF condition.

Fig. 11: User experience feedback provided by the control (C) and counterfactual (CF) groups for criteria listed in Table 5.

for greater interpretative support in CoFact. Specifically, participants would like the system to make suggestions about the kind of judgments the user study asked them to make. We would like to explore automating the calculation and communication of visual suggestions about a feature's importance.

7 CONCLUSION

This paper presented a novel counterfactual approach that reveals the presence of confounding factors during visual analysis, implemented in the CoFact prototype system. CoFact enables users to interactively explore data, apply filter constraints, and analyze the resulting included, excluded, and counterfactual subsets. Visualization of the proposed counterfactual subset, alongside other descriptive information, encourages users to think more critically about feature-to-outcome relationships and the potential of counterfactual possibilities during data exploration.

A controlled user study (n = 30) was conducted to evaluate CoFact, followed by semi-structured interviews about participants' overall experience. Results indicate that the counterfactual visualizations led to improved user inference without complicating the interface significantly. With the counterfactual visualizations, users exhibited greater confidence in strong outcome indicators and lower confidence for weak outcome indicators. A thematic analysis of interviews suggested that participants appreciated the counterfactual approach and would find it useful for data exploration and decision-making. Key areas for future work include (1) greater sophistication and customization in determining counterfactual subsets and (2) calculation and communication of interpretive, actionable counterfactual insights.
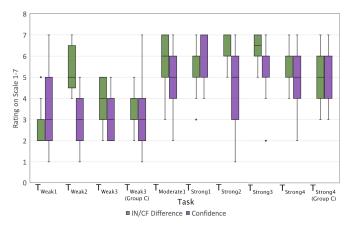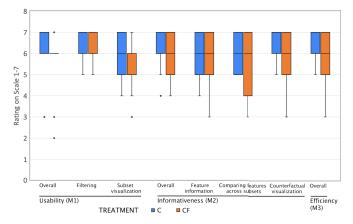
ACKNOWLEDGMENTS

Fig. 10: Difference between included (IN) and counterfactual (CF) subset outcome distributions and corresponding confidence levels.

limited amount of time to be introduced to, digest, and practice the counterfactual approach using the visual system. This likely exacerbated confusion and risked encouraging participants to give quick, non-thorough answers that were truly lower in confidence than was conveyed. Second, participants sometimes found the confidence question (**Q3** in Table 4) ambiguous. Although the term captures what we intended to study, its ambiguity could have affected participants' responses, and we may seek more precise measures in the future.

6.2 Future Work

This work used a simple methodology to calculate the counterfactual subset as described in Section 3.1. Future work should find more sophisticated methods for this task. In addition, future iterations of this prototype could let users customize various aspects of the counterfactual subset calculation, such as size and alternative similarity metrics, e.g., entropy-based measures. Additionally, several participants asked

REFERENCES

[1] E. Artur and R. Minghim. A novel visual approach for enhanced attribute analysis and selection. *Computers & Graphics*, 84:160–172, Nov. 2019. doi: 10.1016/j.cag.2019.08.015

[2] D. Borland, W. Wang, and D. Gotz. Contextual Visualization. *IEEE Computer Graphics and Applications*, 38:17–23, 2018. doi: 10.1109/MCG.2018.2874782

[3] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, Jan. 2006. Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa. doi: 10.1191/1478088706qp063oa

[4] R. M. J. Byrne. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 6276–6282. International Joint Conferences on Artificial Intelligence Organization, Macao, China, Aug. 2019. doi: 10.24963/ijcai.2019/876

[5] M. Chen, A. Trefethen, R. Banares-Alcantara, M. Jirotka, B. Coecke, T. Ertl, and A. Schmidt. From Data Analysis and Visualization to Causality Discovery. *Computer*, 44(10):84–87, Oct. 2011. doi: 10.1109/MC.2011.313

[6] F. Cheng, Y. Ming, and H. Qu. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. Aug. 2020.

[7] S. Cheng and K. Mueller. The Data Context Map: Fusing Data and Attributes into a Unified Display. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):121–130, Jan. 2016. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2015.2467552

[8] H. Chernoff. The Use of Faces to Represent Points in K-Dimensional Space Graphically. *Journal of the American Statistical Association*, 68(342):361–368, 1973. Publisher: [American Statistical Association, Taylor & Francis, Ltd.]. doi: 10.2307/2284077

[9] J. Choo, H. Lee, J. Kihm, and H. Park. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pp. 27–34, Oct. 2010. doi: 10.1109/VAST.2010.5652443

[10] W. S. Cleveland. *Visualizing Data*. Hobart Press, Summit, New Jersey, 1993. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780141212.

[11] P.-E. Danielsson. Euclidean distance mapping. *Computer Graphics and Image Processing*, 14(3):227–248, Nov. 1980. doi: 10.1016/0146-664X(80)90054-4

[12] S. Diehl, F. Beck, and M. Burch. Uncovering Strengths and Weaknesses of Radial Visualizations—an Empirical Approach. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):935–942, Nov. 2010. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2010.209

[13] G. M. Draper, Y. Livnat, and R. F. Riesenfeld. A Survey of Radial Methods for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 15(5):759–776, Sept. 2009. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2009.23

[14] N. Elmqvist, J. Stasko, and P. Tsigas. DataMeadow: A Visual Canvas for Analysis of Large-Scale Multivariate Data. *Information Visualization*, 7(1):18–33, Mar. 2008. Publisher: SAGE Publications. doi: 10.1057/palgrave.ivs.9500170

[15] N. Elmqvist and P. Tsigas. Causality visualization using animated growing polygons. In *Proceedings of the Ninth annual IEEE conference on Information visualization*, INFOVIS'03, pp. 189–196. IEEE Computer Society, USA, Oct. 2003.

[16] N. Elmqvist and P. Tsigas. Growing squares: animated visualization of causal relations. In *Proceedings of the 2003 ACM symposium on Software visualization*, SoftVis '03, pp. 17–ff. Association for Computing Machinery, New York, NY, USA, June 2003. doi: 10.1145/774833.774836

[17] J. D. Evans. *Straightforward Statistics for the Behavioral Sciences*. Brooks/Cole Pub Co, Pacific Grove, Aug. 1995.

[18] M. Ferreira de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):378–394, July 2003. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2003.1207445

[19] A. Ghazimatin, O. Balalau, R. S. Roy, and G. Weikum. PRINCE: Provider-side Interpretability with Counterfactual Explanations in Recommender Systems. Nov. 2019. doi: 10.1145/3336191.3371824

[20] O. Gomez, S. Holter, J. Yuan, and E. Bertini. Vice: Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, p. 531–535. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3377325.3377536

[21] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual Visual Explanations. *arXiv:1904.07451 [cs, stat]*, June 2019. arXiv: 1904.07451.

[22] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley. DNA visual and analytic data mining. In *Proceedings. Visualization '97 (Cat. No. 97CB36155)*, pp. 437–441, Oct. 1997. doi: 10.1109/VISUAL.1997.663916

[23] P. Hoffman, G. Grinstein, and D. Pinkney. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In *Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management*, NPIVM '99, pp. 9–16. Association for Computing Machinery, New York, NY, USA, Nov. 1999. doi: 10.1145/331770.331775

[24] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. Jan. 2018.

[25] D. Hume. *A Treatise of Human Nature*. Oxford University Press, Oxford, 1978.

[26] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. DimStiller: Workflows for dimensional analysis and reduction. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pp. 3–10, Oct. 2010. doi: 10.1109/VAST.2010.5652392

[27] Z. Jin, S. Guo, N. Chen, D. Weiskopf, D. Gotz, and N. Cao. Visual Causality Analysis of Event Sequence Data. Sept. 2020.

[28] B. Johnson and B. Shneiderman. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *Proceeding Visualization '91*, pp. 284–291, Oct. 1991. doi: 10.1109/VISUAL.1991.175815

[29] N. R. Kadaba, P. P. Irani, and J. Leboe. Visualizing Causal Semantics Using Animations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1254–1261, Nov. 2007. doi: 10.1109/TVCG.2007.70528

[30] Kaggle. House prices: Advanced regression techniques, 2018.

[31] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera. Model-Agnostic Counterfactual Explanations for Consequential Decisions. May 2019.

[32] D. Keim. Designing pixel-oriented visualization techniques: theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, Jan. 2000. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/2945.841121

[33] D. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, Jan. 2002. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/2945.981847

[34] D. Keim and H.-P. Kriegel. VisDB: database exploration using multidimensional visualization. *IEEE Computer Graphics and Applications*, 14(5):40–49, Sept. 1994. Conference Name: IEEE Computer Graphics and Applications. doi: 10.1109/38.310723

[35] D. A. Keim. Visual exploration of large data sets. *Communications of the ACM*, 44(8):38–44, Aug. 2001. doi: 10.1145/381641.381656

[36] D. A. Keim, M. C. Hao, U. Dayal, and M. Hsu. Pixel bar charts: a visualization technique for very large multi-attribute data sets. *Information Visualization*, 1(1):20–34, Mar. 2002. doi: 10.1057/palgrave/ivs/9500003

[37] J. Krause, A. Perer, and K. Ng. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 5686–5697. Association for Computing Machinery, New York, NY, USA, May 2016. doi: 10.1145/2858036.2858529

[38] J. LeBlanc, M. Ward, and N. Wittels. Exploring N-dimensional databases. In *Proceedings of the First IEEE Conference on Visualization: Visualization '90*, pp. 230–237, Oct. 1990. doi: 10.1109/VISUAL.1990.146386

[39] D. Lewis. *Counterfactuals*. John Wiley & Sons, May 2013. Google-Books-ID: bCvnk3JMvfAC.

[40] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. Towards Better Analysis of Deep Convolutional Neural Networks. Apr. 2016.

[41] S. Liu, X. Wang, M. Liu, and J. Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48–56, Mar. 2017. doi: 10.1016/j.visinf.2017.01.006

[42] A. Lucic, H. Oosterhuis, H. Haned, and M. de Rijke. Actionable Interpretability through Optimizable Counterfactual Explanations for Tree Ensembles. Nov. 2019.

[43] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere. Explainable Reinforcement Learning Through a Causal Lens. May 2019.

[44] D. Malinsky and D. Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/phc3.12470. doi: 10.1111/phc3.12470

[45] F. J. Massey Jr. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78, Mar. 1951. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1951.10500769. doi: 10.1080/01621459.1951.10500769

[46] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer. Guiding feature subset selection with an interactive visualization. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 111–120, Oct. 2011. doi: 10.1109/VAST.2011.6102448

[47] T. Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. June 2017.

[48] Y. Ming, H. Qu, and E. Bertini. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *arXiv:1807.06228 [cs, stat]*, July 2018. arXiv: 1807.06228.

[49] R. K. Mothilal, A. Sharma, and C. Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. May 2019. doi: 10.1145/3351095.3372850

[50] Pak Chung Wong and J. Thomas. Visual Analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, Sept. 2004. doi: 10.1109/MCG.2004.39

[51] J. Pearl. Causal Diagrams for Empirical Research. *Biometrika*, 82(4):669–688, 1995. Publisher: [Oxford University Press, Biometrika Trust]. doi: 10.2307/2337329

[52] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[53] J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(0):96–146, 2009. doi: 10.1214/09-SS057

[54] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350. ACM, New York NY USA, Feb. 2020. doi: 10.1145/3375627.3375850

[55] ProPublica. Compas scores, may 2016.

[56] J. Raifman, K. Nocka, D. Jones, J. Bor, S. Lipson, J. Jay, and P. Chan. Covid-19 us state policy database, 2020.

[57] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Feb. 2016.

[58] J. Seo and B. Shneiderman. A Rank-by-Feature Framework for Interactive Exploration of Multidimensional Data:. *Information Visualization*, May 2005. Publisher: SAGE PublicationsSage UK: London, England. doi: 10.1057/palgrave.ivs.9500091

[59] S. Shimizu. Lingam: Non-Gaussian Methods for Estimating Causal Structures. *Behaviormetrika*, 41(1):65–98, Jan. 2014. doi: 10.2333/bhmk.41.65

[60] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, Jan. 1992. doi: 10.1145/102377.115768

[61] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pp. 336–343, Sept. 1996. ISSN: 1049-2615. doi: 10.1109/VL.1996.545307

[62] D. G. Simpson. Minimum Hellinger Distance Estimation for the Analysis of Count Data. *Journal of the American Statistical Association*, 82(399):802–807, 1987. doi: 10.2307/2288789

[63] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. July 2019. doi: 10.1109/TVCG.2019.2934629

[64] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000. Publisher: The MIT Press.

[65] P. Spirtes, C. N. Glymour, R. Scheines, P. Spirtes, C. Glymour, and R. Scheines. Causality from probability. In *Conference Proceedings: Advanced Computing for the Social Sciences, Williamsburgh*, 1990.

[66] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush. LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. June 2016.

[67] A. Tatu, G. Albuquerque, M. Eisemann, P. Bak, H. Theisel, M. Magnor, and D. Keim. Automated Analytical Methods to Support Visual Exploration of High-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):584–597, May 2011. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2010.242

[68] T. N. Y. Times. Coronavirus (covid-19) data in the united states: State-level data, August 2020.

[69] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Nov. 2017.

[70] J. Wang and K. Mueller. The Visual Causality Analyst: An Interactive Interface for Causal Reasoning. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):230–239, Jan. 2016. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2015.2467931

[71] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, Jan. 2020. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2019.2934619

[72] H. Wickham. Tidy data. *The Journal of Statistical Software*, 59, 2014.

[73] J. S. Yi, Y. a. Kang, J. Stasko, and J. Jacko. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, Nov. 2007. doi: 10.1109/TVCG.2007.70515