Understanding Data Analysis Workflows on Spreadsheets: Roadblocks and Opportunities

Pingjing Yang
University of Illinois (UIUC)
py2@illinois.edu

Ti-Chung Cheng*
University of Illinois (UIUC)
tcheng10@illinois.edu

Sajjadur Rahman*
University of Illinois (UIUC)
srahman7@illinois.edu

Mangesh Bendre
VISA Research
mbendre@visa.com

Karrie Karahalios
University of Illinois (UIUC)
kkarahal@illinois.edu

Aditya Parameswaran
University of California, Berkeley
adityagp@berkeley.edu

ABSTRACT

Spreadsheets are widely used for data management and analysis by individuals and teams with varying degrees of programming expertise across a spectrum of domains. While several papers have studied the prevalence of errors on spreadsheets and performed ethnographic studies on spreadsheet use, little is known about how spreadsheet users approach and address computational tasks on spreadsheets, especially on relatively large datasets. To understand how users analyze data on spreadsheets, we conducted a study consisting of eight common analytical tasks, with thirty-two participants. Participants developed an execution strategy for each task and then attempted to operationalize this strategy within the spreadsheet system. From examining the study results and transcripts, we identified the successful and unsuccessful strategies participants adopted in addressing the tasks. In general, we find that unsuccessful spreadsheet users had difficulties mapping spreadsheet models to their predetermined execution strategies, comprehending online help documents when trying to learn how to use new formulae, and identifying workarounds when confronted with roadblocks. We identify opportunities to reduce barriers in computational task completion, including improvements to the spreadsheet interface and better training/educational methodologies and tools.

CCS CONCEPTS

• **Human-centered computing → User studies**.

KEYWORDS

Spreadsheets, User study, Task workflow

ACM Reference Format:
Pingjing Yang, Ti-Chung Cheng, Sajjadur Rahman, Mangesh Bendre, Karrie Karahalios, and Aditya Parameswaran. 2020. Understanding Data Analysis Workflows on Spreadsheets: Roadblocks and Opportunities. In *Proceedings of Workshop on Human-In-the-Loop Data Analytics (HILDA'20)*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HILDA'20, 19 June 2020, Portland, OR, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Spreadsheet systems have enjoyed widespread popularity for ad-hoc data management and analysis, across various domains, for over four decades. In fact, roughly one in ten people around the world use spreadsheets [1]. Despite the emergence of a plethora of other BI and data analysis tools, spreadsheets still play a key role in analytics—information workers often shun enterprise solutions with more advanced analytical features for spreadsheets [9, 26].

Even though spreadsheet systems are popular, they can be challenging to use and error-prone, especially for data-intensive tasks, as documented by a recent study of Excel users on Reddit [19]. Prior work has characterized spreadsheet errors [23], identified causes of errors [16], and conducted ethnographic studies on the usage of spreadsheets [8, 18]. However, there is a lack of research on why accomplishing analytical tasks in spreadsheets is challenging and what strategies lead to success or failure. Specifically, comparing and contrasting spreadsheet user performance across tasks, especially via carefully constructed lab studies, is curiously absent in the literature. While spreadsheet developers have suggested best practices for using spreadsheets, these challenges still persist.

Enabling efficient usage of spreadsheets for analytical tasks can impact the user-experience of hundreds of millions of people worldwide. There are a number of research questions that we explore. First, what workflow strategies do people employ when addressing analysis tasks and are there any commonalities among these strategies? Second, what strategies lead to successful completion and what are the causes of failure? Finally, how do people successfully overcome challenges? An in-depth study exploring these questions can help develop and identify best practices for data analysis in spreadsheets, and provide a roadmap for future spreadsheet system development. Answering these questions is of particular interest to the HILDA community, not just because spreadsheet systems are canonical and popular for HILDA, but also because these answers can inform the development of other HILDA tools.

To address these questions, we conducted a user behavioral study in a lab setting with 32 participants to understand spreadsheet workflow challenges. The study required participants to complete multiple analytical tasks of varying degrees of difficulty in Microsoft Excel. We then dissected the study data to identify different user interactions, encoded task completion strategies, and performed both qualitative and quantitative analysis. The results show that creating *tidy* data via approaches like filtering, sorting, and copy-pasting before analysis, led to successful completion of tasks. On the other hand, task failures can occur for a number of reason,

including repetition of incorrect operations due to psychological fixation [5], difficulty in manipulating large datasets due to lack of easy-to-use operations, and incorrect usage of formulae due to lack of familiarity. Finally, people recover from these failures by seeking help from search engines, rather than using the built-in help tools.

We propose a number of approaches to improve data analytics experience with spreadsheets. For example, designing intuitive data manipulation operations to manage and explore large datasets, employing guardrails such as error detection to prevent users from making mistakes, and introducing workflow assistants to guide users with common complex tasks. Our primary contributions are:

- We design and conduct, to the best of our knowledge, the first ever lab-based study targeted at exploring challenges faced while performing analytical tasks on spreadsheets.
- We identify the recipes for success and failures when performing these tasks. Moreover, we also identify a number of strategies that may help users recover from failure scenarios.
- Based on our observations from the study, we propose enhancements for spreadsheet systems that could help users during the data analysis process.

2 RELATED WORK

A number of papers have studied errors in spreadsheets, including categorizing errors [25], linting and error prevention measures [17, 24], and fixing errors once introduced [4]. These papers focus on error detection and prevention and do not study challenges faced by spreadsheet users while performing analytical tasks. Other work has examined why spreadsheets are useful and popular [22]. Nardi et al.'s interview-based study [21] shows how users have difficulties comprehending the relationships between cells as well as the global structure due to fragmented code scattered across cells. Hendry et al. [16] focused on studying formulae, specifically, their usability, comprehensibility, and communicability between users, by interviewing ten spreadsheet users. That study found that even simple formulae were hard to create and understand without extensive knowledge of the data itself. Lawson et al. [18] conducted a survey of experienced and inexperienced spreadsheet users and found substantial differences in their skills and practices. Recent work by Middleton et al. [27] surveyed and interviewed spreadsheet users in a large multinational conglomerate, and discovered several user challenges, such as in reuse and sharing of spreadsheets across users. While these interviews and surveys provided important qualitative insights into spreadsheet use and challenges, a lab-based study allows us to understand and quantify analytical strategies across users for the same set of tasks—specifically what contributes to success and failure, and how spreadsheet systems can be improved to effectively support both novices and experts.

3 STUDY DESIGN

In this study, we explore how spreadsheet users complete analytical tasks on a large dataset during a mixed-method laboratory study. Specifically, we ask the following research questions: a) What workflow patterns do people use when addressing computational tasks in spreadsheets?, b) What strategies enable successful task completion?, c) What approaches lead to failure in task completion?, and d) How do spreadsheet users overcome completion roadblocks?

Participants. A total of 32 people participated in the study—17 females, 15 males, with ages between 19 to 46 ($\mu = 27, \sigma = 7$). Participants had educational backgrounds ranging from high school to Ph.D., and professions ranging from accounting, web design, and economics, to IT and social work. Each participant spent 20 to 40 minutes in our laboratory during the study and received $10/hour. We classified participants into three spreadsheet experience levels via a questionnaire derived from Lawson et al. [18]. Six participants were classified into the *experienced* group, eight into the *inexperienced*, and the other 18 into the *intermediate* group.

Dataset. We used a publicly available Excel spreadsheet dataset from Airbnb [2, 3] for our study. This dataset consisted of 142,042 rows of rental listings with 16 columns describing each listing, such as price, host name, minimum nights, and last review. Data types varied across categories and included text and numeric types. We selected this Airbnb dataset for the following reasons. First, the data was publicly available and was intended for general consumption [3]. Second, the structure of the data was complex enough for us to reasonably expect a variety of sensemaking processes, yet not too complex to confuse participants. Third, we were interested in how people operate on relatively large datasets; the size of the dataset aligned with our motivation. Finally, the dataset was relatable, requiring no specialized knowledge to comprehend it.

Procedures. The study included three stages: planning, execution, and testing. Participants were asked complete 8 tasks. Before the start of the tasks, participants had 5 minutes to familiarize themselves with the dataset. During the planning stage, participants were introduced to a document listing eight tasks. They were required to write down how they would complete the tasks without actually manipulating the spreadsheet data. Then, during the execution stage, participants were asked to implement the approaches they developed on Microsoft Excel. Finally, during the testing stage, participants were allowed to try, test, and revise their approaches, if they encountered challenges during the execution stage. Participants were allowed to use external resources such as online search engines as well as built-in help manuals within Microsoft Excel. We encouraged the participants to talk aloud throughout the study.

Tasks. According to studies on spreadsheet usage [8, 18], the most frequently used data analysis operations are aggregation (AVERAGE, SUM), look up, and search (VLOOKUP, find/replace), data reorganization (sort, filter). The task design was motivated by these frequently used operations. The tasks are listed in Table 1 in the order they were provided to the participants. These tasks were selected to cover a range of common spreadsheet operations [8, 18]. The tasks can be classified into two major categories: *multi-step* and *advanced operation* tasks. Multi-step tasks (tasks 1–5) are those that would require several steps to complete. For example, to find the average price of listings in a city, participants first filter out other cities, then use the AVERAGE formula. This requires at least two steps. For Aggregate (task 1), Search (task 2), Conditional Aggregate (task 3), Filter and Aggregate (task 4), and Aggregate and Search (task 5) tasks, we expected the participants to use find/replace, filtering, sorting, and built-in spreadsheet formulae like AVERAGE and COUNT. For these

five tasks, we gradually increased the complexity by increasing the number of steps required to complete a task. Advanced tasks refer to those that may require fewer steps but specific knowledge about advanced spreadsheet functions to successfully complete. For example, the Lookup task (task 8) requires knowledge of the VLOOKUP function, while the Format task (task 6) requires conditional formatting. We also included a *multi-step open-ended* Explore task (task 7), to evaluate how participants do open-ended exploration—unlike other tasks, this task did not have a "right" answer.

Task and Interview Analysis. We employed a mixed-method approach to analyze the data. First, the study video interviews from the eight tasks were transcribed. Two researchers then used open coding to label categories and sub-categories that emerged from the transcripts and written documents which included the participants' planned approaches, using NVivo [11]. The two researchers then iterated on these themes and used axial coding to identify relationships among the open codes. For each task, we coded which spreadsheet operations were used, what challenges participants encountered, and if and when people sought external help. A codebook was developed from participants' common problem-solving patterns and challenges [7]. We additionally recorded the times taken by the participants for each step along the way, as well as whether the process led to success or failure.

4 RESULTS

In this section, we report and discuss the user activity data collected from the experiments to address our research questions.

4.1 Efforts and Success Rate

We first discuss the overall task completion performance of the participants. We quantify user efforts using the number of attempts made, and the average time spent on each task.

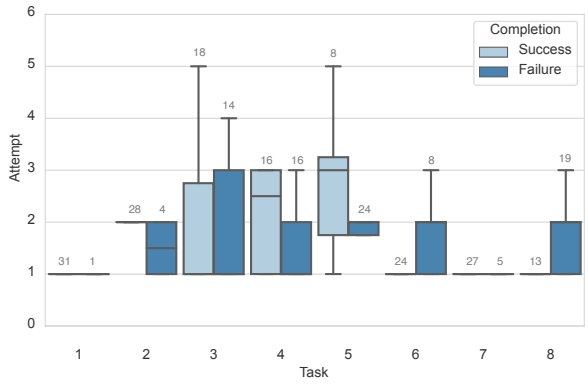

Figure 1: A boxplot showing how the number of attempts differs across success and failure groups for the eight tasks. The numbers on top of whisker lines refers to number of participants that succeeded or failed.

Figure 1 displays the number of attempts participants made during the study. For each task, we constructed two box-plots, one capturing the attempts where participants successfully completed the task (in light blue) and one where they failed (in dark blue). For the Aggregate (1), Search (2), Conditional Aggregate (3), Filter and Aggregate (4), Aggregate and Search (5), and Explore (7) tasks, successful participants made more attempts than those that failed. For example, the Aggregate and Search task (5) required participants to calculate the second largest average city listing price. The eight participants that successfully completed task 5, marked in grey on top of the whisker line, took an average of 2.8 ($\sigma = 1.4$) attempts. For the same task, participants ($N = 24$) who failed only made 2.1 ($\sigma = 0.9$) attempts. However, for advanced operation tasks, such as the Format (6), and Lookup tasks (8), more attempts did not necessarily improve the success rate. In task 8, the average number of attempts for the 13 successful participants is 1.3 ($\sigma = 0.6$), compared to 1.7 ($\sigma = 0.7$) attempts made by the 19 participants who failed. *This suggests that if a task can be decomposed into multiple steps, the success rate increases with additional attempts. However, if a task required advanced knowledge the participant did not possess, making additional attempts did not improve the success rate.*

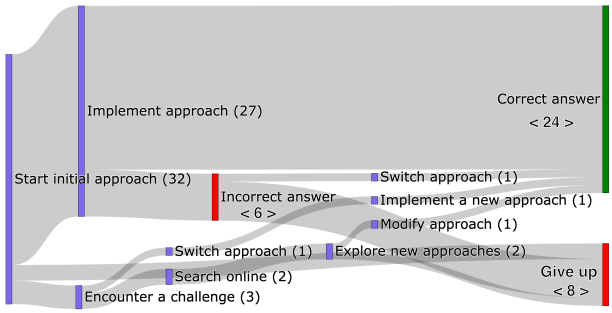

Figure 2: This Sankey diagram shows the progress of participants for the Format task. Each node represents the activity and the number of participants. The edges represent the orders of activities.

Figure 2 shows a fine-grained Sankey diagram summarizing how participants attempted the Format tasks; diagrams for other tasks are similar. Participants start at the "implementing an approach" node, and end with completion nodes such as "Correct answer", "Give up", or "Incorrect answer". Figure 2 shows that 6 participants gave an incorrect answer after performing their planned approaches. Among them, one participant used a different approach to achieve the correct result, while five participants gave up. From studying the Sankey diagrams across tasks, we identified three typical flows for participants when attempting to address tasks: (a) successful submissions–where participants were able to complete a task successfully at the first attempt, (b) refined successful submission–where participants initially failed, but were able to refine their strategies to complete a task. (c) unsuccessful submission–where participants did not recover from a failure.

We summarize the distribution of participants into three categories in Figure 3. Considering the multi-step tasks first (1–5), this chart further demonstrates that that multiple attempts are helpful for multi-step tasks. As we progress in difficulty for the multi-step tasks from 1–5, we end up going from the majority of successful participants having gotten it right in their first attempt, for Aggregate (1) and Search (2), to the majority of successful participants having gotten it right in subsequent attempts for Conditional Aggregate (3), Filter and Aggregate (4), and Aggregate and Search (5). The number of participants who failed also goes up from 1–5, reaching 25 for Aggregate and Search (5), the same number that succeeded on their

Table 1: The description of the computational tasks, listed in the order provided to the participants.

No.	Question	Task Name
1	What is the average(Mean) of the listings' price?	Aggregate
2	What is the price of the listing with the ID equaling 14491416?	Search
3	How many listings in the Harlem neighborhood have a price under 100 dollars?	Conditional Aggregate
4	What ratio of the listings in New York City are frequently rented? A frequently rented listing has less than 60 days of availability in 365 days.	Filter and Aggregate
5	Find the city with the second-highest average listing price across all listings.	Aggregate and Search
6	Highlight all shared room listings via a yellow background. (This requires conditional formatting of such listings.)	Format
7	If I want to travel to San Francisco next month, which listing would you recommend for me to stay in? I want to spend 3 nights there. Give me 2 candidates and reasons.	Explore
8	Use the VLOOKUP formula to return the listing's price by inputting the listing id.	Lookup

first attempt for Aggregate (1). However, the story for advanced operation tasks is not as clear, and debugging via refinement is not as obviously helpful: more participants benefit from multiple attempts for Lookup (8) than for Format (6). Indeed, the syntax for the VLOOKUP operation in the Lookup task is complex, and multiple attempts may be helpful in that case. Most participants ended up getting the open-ended task (7) correct on their first attempt.

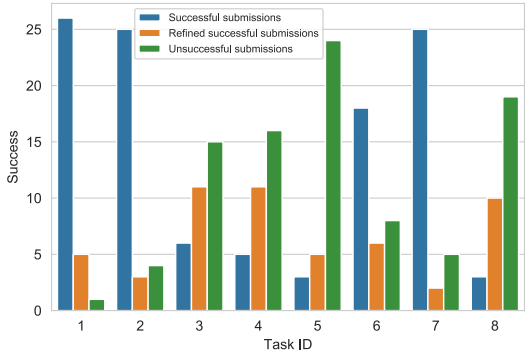

Figure 3: The distribution of cases for 1). Successful submissions, 2). Refined successful submissions, and 3). Unsuccessful submissions categories across eight tasks.

4.2 Successful Task Completion Strategies

Following the approaches created in the planning phase of the study, out of 256 cases (32 participants × 8 tasks), participants successfully implemented their approaches in 111 cases (43.4%) and failed in 145 cases in the subsequent phases. However, for the Aggregate (81.3%, task 1), Search (78.1%, task 2), Format (56.3%, task 6), and Explore (78.1%, task 7) tasks, the participants had a relatively high success rate in implementing their planned approaches. We now analyze the approaches that led to successful task completion.

4.2.1 Pre-processing data. In all of the successful task completion cases, participants adopted various strategies to pre-process the data before performing any analytical operation. P9 reported, "*I guess my first step in all of this would be first to just clean the data because there's a lot of messiness* [...]" Examples of pre-processing operations include copy and pasting a subset of the data, reordering the data rows, and filtering data. We now discuss how these strategies impacted task completion performance.

Copy-pasting a subset: Participants ($N = 10$) copy-pasted a smaller subset of the data onto another spreadsheet sheet to test the validity of their planned approaches—creating smaller subsets made the information more comprehensible. For example, when attempting the Lookup task, P20 copied a small portion of the data to a second sheet and then tested the VLOOKUP formula on the subset to confirm that it worked. This method is an effective workaround to avoid large data processing when working on complicated tasks. After struggling on the raw data sheet, P14 reported, "*Okay so, let me copy this number up here* [a new worksheet]."

Reordering the rows: Before performing a task, participants often ($N = 21$) sort to create a meaningful ordering of the data. For example, P29 started with sorting the data in five of the eight tasks. For the Filter and Aggregate task (task 4), they first sorted the data by the *city* and *availability* columns and then scrolled to locate rows with a price value of 60. Again for the Format task (6), they sorted the data by room type and then highlighted all the shared room rows. By using the sorting operation, participants were able to locate subsets of data quickly. P29 reported, "*so you just sort based on any column, you can look up anything really quickly.*"

Filtering the rows: When attempting a task, participants ($N = 15$) filtered the data by specific column values to hide unnecessary data from the tasks and make the data more perceptually scalable. P16 filtered the data to display listings in the "Harlem" neighborhood for the Conditional Aggregate task (3). Then, they used a combination of the IF and SUM formula to calculate the number of listings that satisfied the availability condition. For the Filter and Aggregate task (4), they first filtered "New York City" in the city column and then applied statistical formulae to calculate the ratio.

4.2.2 Compartmentalizing tasks within separate sheets. Participants ($N = 11$) often made use of multiple spreadsheets within workbook to address different tasks. Each spreadsheet is treated as a separate work-space where participants tested and ran formulae for different tasks. The participants found this strategy helpful for compartmentalizing each task within one sheet, and avoid interference with operations conducted for previous tasks. For example, P29 created different sheets for each task, renaming the sheet-name using the task ID to manage multiple sheets. They only copy-pasted data as necessary into the corresponding sheet. Although we did not ask the participants to reflect on all the approaches after the execution stage, at least one participant volunteered that "*managing different tasks in different sheets will help me to check the correctness of each approach afterward*" (P31).

4.2.3 Preserving the original dataset. Some participants ($N = 5$) attempted each task on the same sheet. After they completed the task, they would undo or remove the operations they had implemented. A clean slate helps participants avoid mutual interference from different tasks. Reversing to a clean slate is similar to switching to

a previous version. Current spreadsheet systems do not explicitly support version control, requiring participants to devise their own mechanisms to "revert" to a clean slate.. They can either maintain a clean version of the original dataset in a separate file or remove all the applied operations after completing each task. We found that participants often chose to use undo to maintain a clean slate.

4.3 Spreadsheet Challenges

There were a significant number of cases (92 out of 256 cases) where participants failed to complete their tasks. Among these cases, 21 were coded as "No idea" in our codebook, indicating that the participants did not even attempt the task. They reported two types of difficulties: a) difficulty in a planning a workflow, and b) unfamiliarity with spreadsheet operations. For example, for the Aggregate and Search (5) task, five participants failed to plan a step-by-step approach to calculate the average price for each city. Several participants ($N = 11$) reported that they "*have no idea what* VLOOKUP *is*" even after reading online tutorials. P25 said "*I just thought it would be easy to use* VLOOKUP*. . . I want to know what value is associated with that* [. . . points to a field in VLOOKUP. . .] *I want to know all the values associated with that number* [. . .] *But I've never really used that before.* " The participants also failed to complete tasks for other reasons. We summarized these challenges below:

4.3.1 Repeating mistakes: psychological fixation. Participants who often failed to complete tasks showed a tendency to reuse their approaches across multiple tasks. Despite the fact that the approaches did not work for more than one task, participants persisted in using them for subsequent tasks. Such behavior may be explained by the psychological fixation phenomenon [5], which refers to people's inclination towards reusing known methods when facing an unknown problem. For example, P27 attempted to use the MATCH formula to address the Conditional Aggregate task (3) even though they did not fully understand when and how it is used. Yet they tried to use the match formula for the Conditional Aggregate (3), Filter and Aggregate (4), and Aggregate and Search (5) tasks. During the study, we encouraged the participants to explore new methods if we found them using the same ineffective approaches repetitively. Unfortunately, we observed that participants were not able to either find a new approach by themselves or by using a search engine. P27 reported, "*This is the only method I know to tackle this . . .*"

4.3.2 Errors when using formulae. We identified three roadblocks participants faced when constructing a formula: errors in a) identifying, b) comprehending, and c) issuing formulae.

Failure in identifying an appropriate formula: Participants ($N = 2$) often could not figure out which formula to use during tasks. P13 reported that "*I am not sure which formula to use, but I want to calculate the second largest value for* [. . .]" P23 reported, "*Yeah I'm kind of frustrated at that one. I would rather load the data into a database or write out actual code for that.*"

Difficulty comprehending formula usage: 25 participants successfully identified an appropriate formula to use from online search results. However, they were not able to apply the formula to their tasks. For example, upon reading the online tutorials on VLOOKUP, P32 failed to understand why and how VLOOKUP would aid in the task of locating the price of a specific listing based on its ID.

Semantic errors with advanced formulae: Participants were often not able to correctly fill in the appropriate arguments for formulae, despite understanding the purpose of the formula as well as its input/output semantics. We found participants especially struggled to fill the parameters for the VLOOKUP formula. Some participants did not understand what the second argument, i.e., the range variable in the VLOOKUP formula, should be. Failure to understand the formula semantics resulted in the participants ($N = 15$) giving up on tasks. After testing out the parameters of the VLOOKUP formula in the function arguments panel in the spreadsheet, P24 reported, "*still there is something wrong* [. . .]"

4.3.3 Scalability-related failures. The scale of the data adversely impacted the success rate of the tasks. Operations that were easy to perform and comprehend in small datasets, were challenging for participants to implement with the dataset we provided. Selecting data is a common operation for almost every task; participants often selected the incorrect data range. For example, while performing the Aggregate and Search (5) task, out of 14 participants who tried to manually select subsets of data by dragging the mouse pointer across several screens, only one succeeded in selecting the appropriate data range in their first attempt. For the Conditional Aggregate (3) and Filter and Aggregate (4) tasks, participants exhibited similar behavior, with only two and four participants succeeding in their first attempt, respectively. Some participants ended up scrolling endlessly—often losing context due to a lack of understanding of the overall structure of the data [28]. Others gave up after spending more than 60 seconds trying to select the dataset. P8 reported the major challenge was "*scrolling all the way through* [. . .] *It's tedious* [. . .] *There should be a fast way, like typing to find the end.*"

One approach for making large datasets easier to work with is to sort it first. Sorting imposes an ordering on the data, and makes it easier to find specific rows. Even with sorted data, however, finding specific rows with desired values within the spreadsheet was challenging. Participants often found it hard to stop at the exact row that contained the target value. For example, some participants ($N = 6$) attempted to find listings from the "Harlem" neighborhood by scrolling to the rows where the column corresponding to the neighborhood started with an "H." They would often inadvertently miss the first few "Harlem" rows. Participants would sometimes remedy this by scrolling up and down repeatedly. As an alternative, some participants switched to using the inbuilt spreadsheet search capabilities to locate the "Harlem" listings.

4.4 Strategies for Recovering from Failures

Overall, 24 participants across 53 separate task instances successfully overcame challenges that arose when implementing their original approaches. We now summarize the recovery strategies.

4.4.1 Iterative refinement of strategies. Out of the 256 (32×8) separate planned approaches across participants and tasks, participants revised 50 of the approaches during the Execution stage. These included both minor ($N = 41$) and major ($N = 9$) revisions. Minor revisions are small adjustments like removing or adding some steps to the original plan, whereas major revisions refers to significant modifications of the initial approaches, such as changing all of the steps. Specifically, eight participants switched from using

an advanced operation to multiple simple operations (in essence, switching from an advanced operation approach to a multi-step operation approach) to derive answers. For example, when P1 could not figure out the right parameters for conditional formatting, they decided to use the highlight feature instead. P16 used a combination of the IF and AVERAGE functions to filter out data satisfying certain conditions after encountering problems with AVERAGEIF. In 41 cases, participants made minor revisions to their approaches. For instance, P19 first scrolled through the data and realized missing filtering criteria. They then adjusted the filtering criteria to include a filter on price to the existing filter on city. P19 shouted: "*It worked!*" when the task was completed. These intermediate steps were not previously planned, but uncovered as the participants worked towards their goals. These examples illustrate the power of iterating on planned strategies when attempting to recover from failures.

4.4.2 Exploration of External Resources. Fourteen participants chose to search online when they encountered challenges while implementing their planned approaches. Among them, five participants revised their planned approaches. For example, P22 noticed that they could not manually list all of the city names to solve the Aggregate and Search (5) task. To overcome this challenge, P22 searched for "*remove redundant cells in excel.*" on a search engine. P22 then revised their plan, extracted the city names, and applied the AVERAGE formula for each of the cities accordingly. In another case, P28 learned about adding one more criterion to the sorting panel while performing sorting, from examining online search results. Participants often followed an ad-hoc error-driven online search strategy to try to address their challenges. Searching during the implementation of approaches was often more effective than searching during the planning stage. For instance, P28 mentioned, "*I don't know the parameters for AVERAGE* ". P28 then searched for "*average function in excel* " online to learn how to fill the appropriate parameters for the AVERAGE formula. Noticing the difficulty in selecting the entire column for the formula, P28 used the search phrase "*Excel how to call the last row*" to learn how to select an entire column for the AVERAGE formula. P28 was able to complete the Aggregate (1) task.

These search processes were iterative; when participants did not find what they were looking for, they often refined their search phrases. For example, P28 started by searching online for "*double search in excel*" to learn how to filter data using two criteria at once. After exploring the search results, P28 was able to revise their search query to "*how to select key values in excel*", and learned how to use sorting by two separate criteria to locate desired values.

5 DISCUSSION

We now discuss how to further improve spreadsheet user experience based on takeaways from our study.

Providing Guardrails to Handle Failures. Not all of the challenges encountered during the study were related to participants' spreadsheet expertise. Some errors were due to the lack of robustness of spreadsheet systems, such as a system crash (see Section 4.3.1 and 4.3.2). While a sequence of actions may result in a spreadsheet crashing, participants ($N = 3$) often repeated those actions upon restarting the system, resulting in the same outcome. Early detection of whether the system or the user is to blame for a failure can help avoid wasted user effort. One way to automate the detection process is to analyze error reports submitted by users during a system crash and identify common patterns or behaviors that cause such failures. Similarly, detecting whether a user action is consistent with the previous sequence of actions can be useful. For example, for the Filter and Aggregate (4) task, two participants issued a COUNTIF formula on the data filtered by city. This approach provided incorrect results, as Excel applies the formula on the entire spreadsheet range. Providing prompts or warnings explaining the potentially incorrect usage can help users avoid such mistakes. Semantic errors while issuing advanced formulae can also be prevented by such preemptive detection mechanisms.

Automating Spreadsheet Learning. One of the challenges associated with spreadsheet systems is learning to use the many available features—most of which are quite complex to master (see Section 4.3.2). While our computing environments have undergone significant changes over the past few decades, the help manuals or tutorials for spreadsheet systems has not changed substantially. Our study showed that to learn complex features like Pivot Table or VLOOKUP, participants often resorted to searching online, watching video tutorials, or exploring Excel help communities. However, the search process is manual, cumbersome, and often results in participants giving up on a task. One way to address this issue is to provide automated guidance or supervision for users as they use complex spreadsheet features. Similar ideas have been adopted in other domains. For example, CommunityCommands [20] recommends learning material by collecting and analyzing software usage data from thousands of Autodesk users, and then generating personalized command recommendations or recipes. Another approach can be to provide users with rapid, contextual, and within-spreadsheet access video clips, to help them understand how to use the associated features, similar to ToolClip [14].

Supporting Data Exploration at Scale. Our study reveals that manipulating and exploring spreadsheet data can be challenging. For example, participants failed to select appropriate data ranges for the Aggregate and Search task (see Section 4.3.3). In all of these cases, the desired subsets of data spanned thousands of rows making it cumbersome and error-prone to manually manipulate the data, such as by dragging a cursor across multiple screens. These challenges were evident even in earlier spreadsheet systems that spanned only a handful of rows, as shown in prior work [10, 21, 28]. In recent years, spreadsheet systems have stretched to support increasingly large datasets: 10s of billions of cells for Microsoft Excel [12] and five million cells [13] for web-based Google Sheets. Therefore, the challenges related to data exploration and manipulation have been magnified more due to increasing data sizes. To address these challenges, spreadsheet systems should provide a more intuitive interface that enables users to interact with large datasets efficiently and effectively. One approach would be to integrate an overview of the overall structure of the data within the spreadsheet [15]. By linking the overview with the underlying data, spreadsheet users can manipulate large collections of data without having to tediously scroll or drag the mouse pointer. Our recent attempt at integrating an overview plug-in for spreadsheets [6] is one step in this direction, but more work remains to be done.

REFERENCES

[1] [n.d.]. How finance leaders can drive performance. bit.ly/ms_excel_finance.
[2] [n.d.]. The inside airbnb about page. http://insideairbnb.com/about.html.
[3] [n.d.]. The inside airbnb dataset. http://insideairbnb.com/get-the-data.html.
[4] Yanif Ahmad, Tudor Antoniu, Sharon Goldwater, and Shriram Krishnamurthi. 2003. A type system for statically detecting spreadsheet errors. In *18th IEEE International Conference on Automated Software Engineering, 2003. Proceedings.* IEEE, 174–183.
[5] Salman Akhtar. 2018. *Comprehensive dictionary of psychoanalysis.* Routledge.
[6] Mangesh Bendre, Tana Wattanawaroon, Sajjadur Rahman, Kelly Mack, Yuyang Liu, Shichu Zhu, Yu Lu, Ping-Jing Yang, Xinyan Zhou, Kevin Chen-Chuan Chang, et al. 2019. Faster, higher, stronger: Redesigning spreadsheets for scale. In *2019 IEEE 35th International Conference on Data Engineering (ICDE).* IEEE, 1972–1975.
[7] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development.* sage.
[8] David A Bradbard, Charles Alvis, and Richard Morris. 2014. Spreadsheet usage by management accountants: An exploratory study. *Journal of Accounting Education* 32, 4 (2014), 24–30.
[9] Yolande E Chan and Veda C Storey. 1996. The use of spreadsheets in organizations: Determinants and consequences. *Information & Management* 31, 3 (1996), 119–134.
[10] Andy Cockburn, Amy Karlson, and Benjamin B Bederson. 2009. A review of overview+ detail, zooming, and focus+ context interfaces. *ACM Computing Surveys (CSUR)* 41, 1 (2009), 2.
[11] Bengt Edhlund and Allan Mcdougall. 2016. *Nvivo 11 Essentials.* Lulu.com.
[12] Excel scale [n.d.]. Excel limit. https://support.office.com/en-us/article/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3.
[13] Google Sheets scale [n.d.]. Google sheets limit. https://bit.ly/gs_limit.
[14] Tovi Grossman and George Fitzmaurice. 2010. ToolClips: an investigation of contextual video assistance for functionality understanding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 1515–1524.
[15] Jonathan Grudin. 2001. Partitioning digital worlds: focal and peripheral awareness in multiple monitor use. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 458–465.
[16] D.G. Hendry and T.R.G. Green. 1994. Creating, Comprehending and Explaining Spreadsheets. *Int. J. Hum.-Comput. Stud.* 40, 6 (June 1994), 1033–1065. https://doi.org/10.1006/ijhc.1994.1047
[17] Daniel Kulesz, Verena Käfer, and Stefan Wagner. 2018. Spreadsheet guardian: An approach to protecting semantic correctness throughout the evolution of spreadsheets. *Journal of Software: Evolution and Process* 30, 9 (2018), e1934.
[18] Barry R. Lawson, Kenneth R. Baker, Stephen G. Powell, and Lynn Foster-Johnson. 2009. A comparison of spreadsheet users with different levels of experience. *Omega* 37, 3 (2009), 579–590. https://doi.org/10.1016/j.omega.2007.12.004
[19] Kelly Mack, John Lee, Kevin Chang, Karrie Karahalios, and Aditya Parameswaran. 2018. Characterizing Scalability Issues in Spreadsheet Software using Online Forums. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, CS04.
[20] Justin Matejka, Wei Li, Tovi Grossman, and George Fitzmaurice. 2009. CommunityCommands: command recommendations for software applications. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology.* 193–202.
[21] Bonnie A Nardi and James R Miller. 1990. *The spreadsheet interface: A basis for end user programming.* Hewlett-Packard Laboratories.
[22] Donald A Norman. 2002. The Design of Everyday Things. (2002).
[23] Raymond R Panko. 1998. What we know about spreadsheet errors. *Journal of Organizational and End User Computing (JOEUC)* 10, 2 (1998), 15–21.
[24] Raymond R Panko. 2008. Spreadsheet errors: What we know. what we think we can do. *arXiv preprint arXiv:0802.3457* (2008).
[25] Stephen G Powell, Kenneth R Baker, and Barry Lawson. 2008. A critical review of the literature on spreadsheet errors. *Decision Support Systems* 46, 1 (2008), 128–138.
[26] Neil Raden. 2005. Shedding light on shadow IT: Is Excel running your business. *DSSResources.com* 26 (2005).
[27] Justin Smith, Justin A Middleton, and Nicholas A Kraft. 2017. Spreadsheet practices and challenges in a large multinational conglomerate. In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC).* IEEE, 155–163.
[28] Jennifer Watts-Perotti and David D Woods. 1999. How experienced users avoid getting lost in large display networks. *International Journal of Human-Computer Interaction* 11, 4 (1999), 269–299.