Rethinking the First Look at Data by Framing It

Sara Alspaugh
Anna Swigart
Ian MacFarland
Randy H. Katz
Marti Hearst

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2015-208
http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-208.html

November 2, 2015

Acknowledgement

This research is supported in part by NSF CISE Expeditions Award CCF-1139158, DOE Award SN10040 DE-SC0012463, and DARPA XData Award FA8750-12-2-0331, and gifts from Amazon Web Services, Google, IBM, SAP, The Thomas and Stacey Siebel Foundation, Adatao, Adobe, Apple, Inc., Blue Goji, Bosch, Cisco, Cray, Cloudera, EMC2, Ericsson, Facebook, Fujitsu, Guavus, HP, Huawei, Informatica, Intel, Microsoft, NetApp, Pivotal, Samsung, Schlumberger, Splunk, Virdata and VMware.

Rethinking the First Look at Data by Framing It

S. Alspaugh, A. Swigart, I. MacFarland, R. Katz and M. Hearst
University of California, Berkeley

Abstract—**Full-featured data analysis tools provide users a wide variety of ways to transform and visualize their data; ironically, this abundance can be as much hindrance as help in the initial stage of data exploration. In these stages, the critical question is often not "what steps must I take to visualize this data?" but rather "what is this data and what can it tell me?" This mismatch leads to several intertwined challenges. It's difficult to get a mental picture of the data without first visualizing it, but it's hard to identify the appropriate way to visualize the data without first having a mental picture of it. Moreover, it's all too easy for an intriguing data point to pique a researcher's interest and distract them from their current task. This difficult-to-navigate and distraction-rich environment can easily hide faulty assumptions from notice until they botch the analysis later down the line. Together these problems can send the analyst tumbling down a rabbit-hole of progressively deeper and sometimes misguided analysis, while the remainder of the data landscape lies uncharted. We investigate whether we can address these problems through a set of interface features that could easily be incorporated into current visual analytics tools. We built a prototype implementation of these features called DataFramer. Preliminary assessment via a study with 29 participants suggests the approach of examining data and stating questions before exploring the data is promising. We present a taxonomy of exploratory analysis statements and errors, as well as qualitative observations about how participants posed questions for exploring data using different tools.**

Keywords-**data analysis ; data visualization**

I. Introduction

Exploratory data analysis is key for making effective use of big data. It is undertaken at the beginning of an analysis to familiarize oneself with a dataset. Typical goals include suggesting hypotheses, assessing assumptions, and supporting future analyses decisions. It involves heavy use of visualization and descriptive techniques like summary statistics, histograms, and clustering. It is surprisingly difficult to find a precise definition of the exploratory data analysis even in the writings of its famous champion, John Tukey, For example, Tukey writes:

> "If we need a short suggestion of what exploratory data analysis is, I would suggest that: 1. it is an attitude, AND 2. a flexibility, AND 3. some graph paper (or transparencies, or both)." [1]

This reluctance to be more specific likely reflects a desire to emphasize the importance of flexible thinking, and to balance against what was, at the time, a relative overemphasis on prescriptive and confirmatory statistics. Today, despite being a necessary phase of any analysis, it remains a nebulous art, defined by an attitude and a collection of techniques, rather than a systematic methodology. We argue that because of this, the exploratory data analysis process is not well-supported by many tools.

When we have observed inexpert practitioners[1] perform exploratory data analysis using popular visual analytics tools, such as Tableau, we have found that lack of explicit support for exploratory data analysis leads to several challenges:

Lack of support for improvised workflows: The exploration process is inherently open-ended and emphasizes hypothesis generation. However, many tools don't help users articulate and track analysis goals. Instead, users either laboriously track goals in a separate tool, attempt to remember them, or simply neglect to consciously articulate questions at all, and instead fall into a pattern of casting about aimlessly, The resultant confusions, interruptions, and cognitive load degrade the exploration experience.

Unnecessary tedium due to repetition: Tasks such as examining histograms and scatter plots are common to most exploratory analyses. But many tools make it tedious to generate such plots for all desired subsets of the data. Such work could be easily automated and presented upon request.

Tension between unfamiliarity and specification: It's difficult to get a mental picture of the data without first visualizing it, but it's hard to identify the appropriate way to visualize the data without first having a mental picture of it. To generate visualizations, many tools require users to specify the low-level sequence of transformations they need to achieve their desired effect. These can be hard to figure out without a prior familiarity with the domain, resulting in frustration.

Hidden erroneous assumptions: The above challenges lead many users to inadvertently pursue dead-end lines of analysis based on false yet often easily checked assumptions about the data, typically concerning its extent or completeness. Since exploratory analysis is meant to isolate "patterns and features of the data and reveals these forcefully to the analyst," tools should make it difficult for the user to avoid checking some of these typical erroneous assumptions [3]. These problems prevent users from "getting in the flow" by straining memory and necessitating interruptions [2].

[1] We conducted pilot studies in a research seminar. These practitioners were *intermediate users* as defined by Bederson [2].

Given the challenges resulting from this uncertainty about what constitutes a good exploratory data analysis process, our goals are to

- Investigate details of the beginning stages of exploratory data analysis behavior, and
- Mitigate some of the problems described above.

As a vehicle for these investigations, we created a prototype tool, DataFramer, aimed at facilitating the earliest stages of data exploration. We then compared DataFramer to other tools to try to understand how the differences in design impact exploration behavior.

Our overarching hypothesis is that analysts will make better decisions and fewer errors if they first think about the questions that are applicable and study the form of their data before diving into the analysis. We further hypothesize that a visual overview of a dataset using a tool like DataFramer would be a complement to standard tools. We conducted a study in which we asked data analysts to examine datasets and posed questions. We compared the quality and accuracy of the questions and observations they produced. While we did not find a significant difference between tools in the number of errors, we did find that participants were able to use the question formation features of DataFramer, and we uncovered a rich set of patterns in the errors produced using both tools that instruct the design of the next round of features. Overall, the approach of examining data and stating questions before exploring the data seems promising.

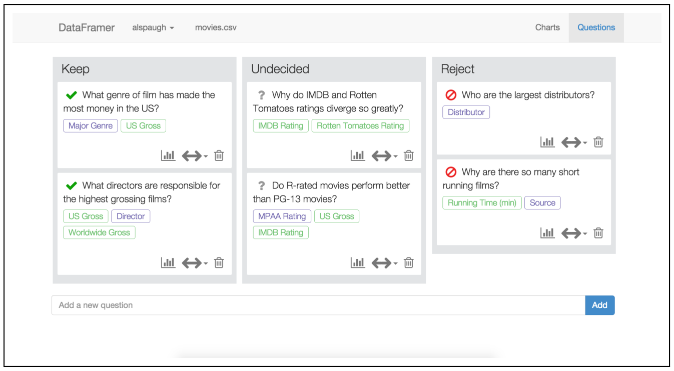
II. Related Work

Exploratory data analysis tools: In general, there are more tools for visualizing and analyzing data than we have room to cite here; instead we focus on those that have been most influential to us [4]. Tableau is descended from earlier research in exploratory data analysis and visualization, including Polaris and Automatic Presentation Tool [5]–[7]. DataWrangler is a tool for facilitating data cleaning, in particular, data reformatting [8]. In a similar vein, Profiler is a tool for data quality assessment, for instance, finding missing values, outliers, and misspellings [9]. Like our prototype, these tools try to help users understand their data, though they focus on data cleaning, rather than on data exploration. In a spirit closely aligned with our goal, Perer and Shneiderman propose a framework, called SYF (Systematic, Yet Flexible), for guiding users through exploration of social networks [10]. The SYF framework can be implemented in data analysis tools to provide an overview of recommended analysis steps, suggest unexplored states, and allow users to annotate and share a record of their activities. We were also inspired by related guides on how to approach a dataset for the first time [11]–[13]. In his paper on interfaces for staying in the flow, Bederson discusses user types and how interfaces can best support their work without distracting, which is an important goal of ours as well [2].

Reducing unnecessary tedium: The Chimera system implemented five techniques for reducing repetition in graphical editing by automating repetitive tasks [14]. This work focuses on graphical editing broadly construed, not on creating data visualizations, so the techniques do not overlap with our approach. However, it would be useful to incorporate them into exploratory data analysis tools. Other work has focused on making visualizations easier to create by allowing users to specify them via natural language [15]. There is a large body of work on automatically generating visualizations, in some cases for exploring data [7], [16]–[24]. There is additional research on automatically generating data transformations and even custom interfaces to carry out specialized visualization tasks [25], [26]. Currently, our prototype automatically generates simple overviews of each column, but future work could incorporate automatic generation of more complex transformations and visualizations.

Analysis workflow support: HARVEST is a prototype visual analytics system developed designed to support the provenance of insights generated by users [27], [28]. It defines a set of semantics-based interaction primitives, called actions, like filter, sort, and zoom, and exposes to the user their history of actions. It uses this history to make context-driven visualization recommendations to help users find appropriate visualizations for their task. Later research built upon this work by creating a more full-featured tool called Smarter Decisions [29]. These ideas are both related to work on supporting graphical histories in Tableau, though to our knowledge, this technology is not present in the current version of Tableau [30]. There have been many other papers published on graphical histories, in addition to the above work [31]–[34]. As we explain in Section III, DataFramer differs from this body of work in that it supports analysis workflows by explicitly prompting the user to articulate potential analysis questions, rather than just tracking user actions and exposing those actions to the user.

Understanding analysis behavior: Recently, Yang et al. undertook a study to understand how people comprehend composite visualizations in different situations [35]. Our evaluation approach was particularly influenced by their approach to coding their results. Others have investigated how individual characteristics such as perceptual speed and verbal working memory influence the interpretation of visualizations [36]. Some researchers have conducted observational studies on how different classes of users, such as designers, use visualization tools [37]. Still others have conducted case studies of exploratory data analysis [38]–[42]. We were more influenced by our own observations than these case studies, as these tended to be particular to specific visual analytics tools under development by the researchers. Researchers have also proposed models of interaction for exploration and visualization [43]–[45]. We leave the incorporation of such models into exploratory data



✓ What genre of film has made the most money in the US?

✓ What directors are responsible for the highest grossing films?

? Why do IMDB and Rotten Tomatoes ratings diverge so greatly?

? Do R-rated movies perform better than PG-13 movies?

⊘ Who are the largest distributors?

⊘ Why are there so many short running films?

Figure 1. DataFramer encourages users to generate, reflect upon, and organize questions about the dataset under examination.

analysis tool design to future research.

III. Research Questions and Design Goals

As described in Section I, our goals are to

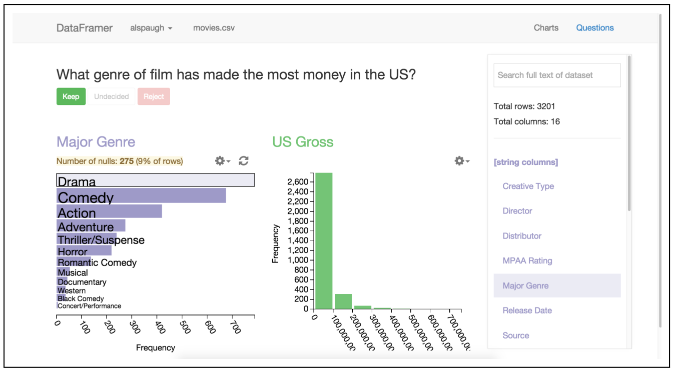
- Investigate details of the beginning stages of exploratory data analysis behavior, and
- Mitigate some of the common exploratory data analysis problems we observed.

To these ends, we created a prototype named DataFramer. DataFramer currently supports exploration of tabular data. Its key features, selected to address the challenges described previously, are as follows:

Question-driven workflow: Expert R package developer Hadley Wickham recently observed, "A good data scientist will help the real domain experts *refine and frame their questions*[2] in a helpful way. Unfortunately I don't know of any good resources for learning how to ask questions." [46] To help users learn to ask questions, DataFramer encourages users to generate, reflect upon, and organize potential lines of inquiry. After uploading data, the user is directed to the questions page (Figure 1). This presents an initially empty Trello-like set of lists for grouping questions into.[3] When the user inputs a question, the question text is placed onto a card. Each card can be moved from one list to the other. This helps the user develop a plan for both what to analyze and what *not* to analyze. By default, DataFramer provides three lists into which questions can be organized:

- **Keep:** questions to pursue in later analyses,
- **Undecided:** questions that are not yet sorted, and
- **Reject:** discarded questions (invalid or unanswerable).

A user may discard a question for any reason. A good reason would be if the question was based on a faulty assumption; for example, the assumption that some piece of information is present in the dataset that actually is not. This process reduces the difficulty of deeper exploration by creating a persistent reference document to support subsequent work.



What genre of film has made the most money in the US?

Figure 2. For any of the questions they pose, users can navigate to a details page. Here they can choose columns to associate with that question. In the future, users will be able to combine selected columns into compound charts that show the relationships between columns.

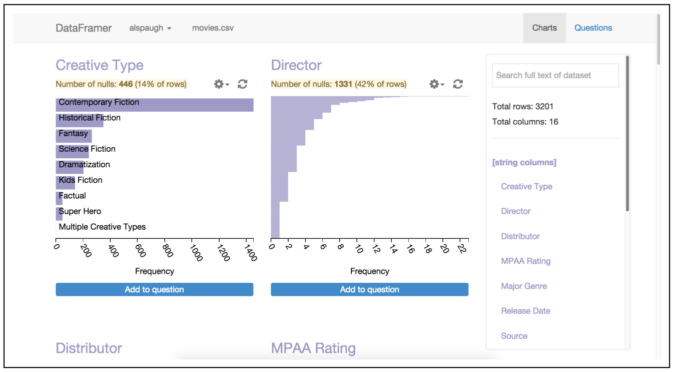


Figure 3. DataFramer automatically generates a visual overview of each column in the dataset, its datatype, and the distribution of values in that column. This lets users quickly survey the fundamental properties of the dataset while avoiding the tedium of creating such visualizations themselves, which allows them focus on asking questions and identifying potential invalid assumptions.

As future work, we plan to enable other ways of organizing questions.

Explicit links between data and questions: DataFramer encourages users to identify which subset of the data they need to answer each question. To note this, users can associate, or tag, questions with the relevant data columns. The associated columns are indicated by a colored tag on the question card. The question card also links to a workspace for that question (Figure 2). This page shows the question along with the overview charts of the columns that the user has already associated with this question. These associations are stored as part of the user's analysis plan. On this page, the user can select and de-select additional columns to associate with the question, which changes the set of displayed charts. As future work, we plan to allow the user to combine selected columns into a compound chart view that displays the relationship among the columns. The question-oriented workspace is designed to help users spot potential problems with their analysis plan, such as incorrect assumptions about the contents of their datasets, with fewer distractions.

[2] Emphasis added.

[3] www.trello.com

Auto-generation of charts: It is difficult to generate questions about a dataset without knowing much about it. Thus, the DataFramer encourages users to begin their exploration by looking at the charts page (Figure 3). There, the user is presented with an overview chart for each of the columns in the dataset. Each chart is a form of frequency distribution whose exact form depends on the corresponding column datatype.[4] Datatypes are automatically detected as part of the upload process and can be manually corrected. Because understanding the types of data present in a dataset is one of the critical goals of data exploration, the interface places emphasis on making the distinctions between these types clear; they are dual-coded with both different colors and different forms for their overview charts. These charts are useful for identifying common gotchas, such as the presence of outliers, or the use of numbers as unique IDs or categories. This frees users from the distracting tedium of repeatedly creating simple charts and lets them immediately focus on the data at a high level. Browsing the charts is likely to prompt many questions about specific columns in the dataset. Users can use the *Add to question* button to associate the given column with a question they have already posed, or to compose a new question to associate the column with. This brings up a modal where users can enter in a new question or check a box next to each question the column should be associated with.

Intentionally limited functionality: So as not to distract users with the ability to transform and visualize data in arbitrarily complex way, DataFramer provides a focused interface targeted at the goal of identifying simple types of assumptions and asking questions. Users do not construct charts with DataFramer; rather, DataFramer automatically creates a summary visualization based on the column datatype. By accepting this constraint at the very outset of the exploration process, users avoid the temptation of diving into rabbit-holes before developing a broad familiarity with their data.

DataFramer differs from prior work on supporting analysis workflows by explicitly prompting the user to articulate potential analyses actions, rather than just tracking the actions taken and exposing them to the user. In DataFramer, users can organize and refer back to the set of questions they intend to pursue in later analyses, but not the actions they took, which is less important because DataFramer intentionally only supports examining univariate distributions, for the time being. In contrast, in the work described in Section II, users can refer to the actions they took during the course of an analysis, but these actions do not necessarily correspond to coherent lines of inquiry, and may be difficult to use to plan out future work.

Name	Description
on-time performance	flight schedules and delays (spans one month)
wildlife strike	collisions of planes with wildlife (spans years)

Table I
PUBLICLY AVAILABLE FAA DATASETS USED IN OUR STUDY

DataFramer is a JavaScript web application that is also easy to deploy locally. Datasets and questions are stored in a database, and each has its own unique, persistent URL, allowing users to bookmark, share, and revisit any content they have created. The code is open-source and can be found at https://github.com/macfarlandian/DataFramer.

IV. Study design

Quantifying exploratory analysis behavior is difficult. Rather than observing unbounded and open-ended exploration sessions, we asked participants to spend twenty minutes exploring some data using an assigned tool. We compared DataFramer to two other types of tools: a typical spreadsheet application (Excel), and a typical visual analytics application (Tableau). We then had participants summarize, in a short memo, what analyses they would propose conducting next. We used these memos as proxies for assessing the analyses users might have performed in a full exploration session.

We are interested in evidence that the challenges identified in Section I are mitigated by DataFramer. In this study, we focus on how often participants make erroneous assumptions about the data that could affect later analyses.

A. Task description

We asked participants to imagine they are a data scientist for a major consulting firm who have been told to spend twenty minutes exploring a sample of data using an assigned tool. They were then to write a short preliminary memo detailing challenges and opportunities for analysis of that data. We provided two datasets, one for each task (Table I).

We assigned participants to groups that dictated which tool and which dataset they used for each task (Table II). If the participants were assigned to use DataFramer for the task, we provided a walk-through before the task began. We asked participants to perform all tasks on their personal laptops and use their preferred text editor to write the memo. For each task, we gave participants a short README that listed a brief description of each column in the dataset.

Participants began the first task, exploring the assigned dataset with the assigned tool. After twenty minutes, we asked them to stop and take a short survey to submit their memo. The survey also asked participants:

- what tools they use to analyze data,
- how much training they have in analyzing data, and
- how many hours per week last year they analyzed data.

They then repeated the task using the other dataset and tool under comparison.

[4] DataFramer currently supports five datatypes: `string` (more broadly, categorical or factor variables), `integer`, `float`, `date`, and `time`. It uses Meteor, a framework that runs on top of NodeJS (runtime environment) and MongoDB (data store) and JavaScript page templates based on AngularJS, styled with Bootstrap, and charts using d3.js.

Task 1 Tool	Task 2 Tool	Task 1 Dataset	Participants
spreadsheet	dataframer	on-time performance	9
spreadsheet	dataframer	wildlife strike	6
dataframer	spreadsheet	on-time performance	4
dataframer	spreadsheet	wildlife strike	5
tableau	dataframer	on-time performance	3
tableau	dataframer	wildlife strike	2
dataframer	tableau	on-time performance	0
dataframer	tableau	wildlife strike	0

Table II
NUMBER OF PARTICIPANTS IN EACH STUDY CONDITION (29 TOTAL)

Category	Description
header	A statement made to indicate memo structure
question	A query about the data
hypothesis	A conjecture or theory about the data
plan	A proposed course of analysis action
action	A description of an analysis step took during the task
observation	An interpretation or inference of the data
metacognition	A description of the thought process applied
context	A fact external to the data or an analysis implication
data complaint	Dissatisfaction with some aspect of the data
tool complaint	Dissatisfaction with some aspect of the tool used
tool compliment	Praise for some aspect of the tool used

Table III
THE TYPES OF STATEMENTS MADE BY PARTICIPANTS IN THEIR MEMOS

B. Data collected

We recruited 33 participants, gathering two memos from each. Memo length did not vary much —participants had been instructed to write approximately ten sentences.[5] We excluded data from four participants from analysis: two used the wrong dataset for one of the tasks, and two others used their memo primarily to provide feedback about the tool.

C. Analysis method

We split each memo into phrases and categorized each phrase by type (Table III). We analyzed every question, hypothesis, plan, and observation and identified a set of commonly occurring errors (Section V). Two members of our team assessed the phrases independently, then compared results to check for agreement. In cases of disagreement, we discussed until consensus was reached. We describe our observations in the next section. We intend this to be a descriptive, exploratory study to inform future study designs.

V. OBSERVATIONS

A. Participant analysis background

The majority of participants (15) said analysis was currently or ever had been a small part of their job or school studies (Figure 4(a)). The majority of participants (19) reported spending an average of four hours or less on data analysis per week over the past year (Figure 4(b)). The number of hours participants report spending analyzing data is closely related to how much analysis training they report having had (Figure 4(c)). The majority of participants report using scripting languages (Python, MATLAB, R) and spreadsheets to analyze data (Figure 4(d)). After that there is a steep drop-off in the number of participants who report using other tools, like database query languages (SQL), cluster computing frameworks (Spark, Hadoop), statistical software (SPSS), command-line utilities (`awk`, `grep`, `sort`), and visual analytics tools (Tableau, Omniture). In most cases, the amount of time users of those tools reported spending analyzing data per week was about the same regardless of whether they used the tool or not. Apart from tools that only one participant reported using, tool usage appears to be somewhat evenly spread out among analysis training levels.

B. Types of statements

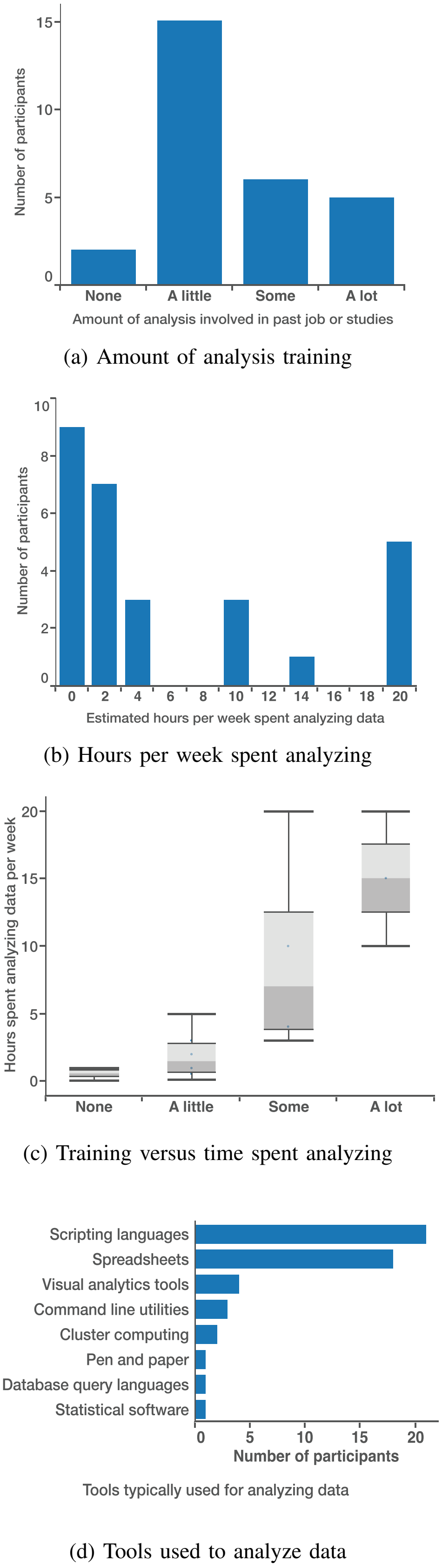
As described in Section IV-C, we split each participant memo into phrases. We then categorized each phrase (Table III). Broadly speaking, there are two statement types: *taken:* those focusing on the analysis that was done, and *planned:* those focusing on the analysis that could be done in the future. The taken category includes action, observation, metacognition, and context, as well as phrases about the data and tool. The planned category includes question, hypothesis, and plan. Below we analyze the errors participants made (taken) that would lead to errors in future analysis (planned).

Figure 5 summarizes the statement frequencies. The most common type of phrase is observation, followed by question, then plan. Participants tended to pose slightly more questions in their memos when using DataFramer as opposed to other tools (Figure 6). In Tableau, the tendency was for participants to focus on describing the actions they took while exploring, posing comparatively few questions or hypotheses. This behavior could be reflective of the greater emphasis in Tableau on constructing visualizations, as opposed to creating calculations and inspecting tables (spreadsheets) or exploring individual columns and asking questions (DataFramer).

C. Data and logic errors

We selected the two types of errors that we could most objectively identify: **data errors** and **logic errors**. Data errors happen because the participant does not understand the data, even though they should be able to check their understanding by examining the data available to them. Logic errors primarily occur when the participant does not understand the implications of applying certain transformations to the data, in terms of the quantity that will result from applying the transformation. We developed a rubric for checking for these errors, which we applied to the subset of planned statements (questions, hypotheses, and plans) and

[5] The median number of sentences per memo was nine (mean: 9.71 and standard deviation: 4.97). The median number of words was 193 (mean: 191.89 and standard deviation: 62.19)



(a) Amount of analysis training

(b) Hours per week spent analyzing

(c) Training versus time spent analyzing

(d) Tools used to analyze data

Figure 4. Participant responses to questions about analysis experience

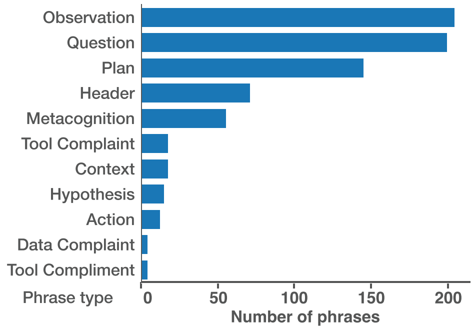


Figure 5. Counts of statements made in memos written by participants; they most frequently contained observations, questions, and plans.

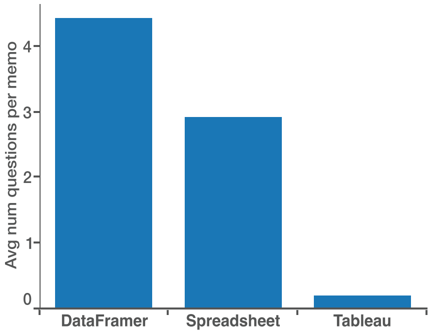


Figure 6. Participants posed more questions in their memos when using DataFramer than when using other tools. Participants using Tableau often posed no questions, focusing instead on their actions and observations.

taken statements (observations) that could be said to be true or false (i.e., did not represent a subjective opinion).

For planned statements, the errors are defined as:

Data error: an assumption about the data that is false

Logic error: flawed reasoning about an analysis plan

For taken statements, the errors are defined as:

Data error: a mistaken calculation or lookup

Logic error: an incorrect or unsound inference

We describe our results in Table IV. We observed that participants made more data errors on average when using spreadsheets than they did with DataFramer though they made the least in Tableau. This pattern held across both datasets and tasks, though no participants used Tableau for the second task. Participants made similar numbers of data errors across tasks, and datasets. Logic errors were similarly close across task, datasets, and tool. Neither the number of data errors or logic errors seemed to have much relationship to the amount of analysis experience reported.

The rest of this subsection illustrates the rubric used to count these data and logic errors. In parentheses are shown the tool used when the example error was made.

1) Data errors in questions, hypotheses, and plans: We found two cases.

• Assuming information is present in the dataset that is not, or assuming some part of the data is of a form that it is not in:

Errors	Tool			Task		Dataset	
data	D	S	T	1	2	on-time	wildlife
average	1.17	2.21	0.80	1.45	1.69	1.62	1.52
std dev	1.44	2.67	0.84	1.62	2.45	2.29	1.84
logic	D	S	T	1	2	on-time	wildlife
average	1.24	1.08	1.40	1.03	1.34	1.07	1.31
std dev	1.46	0.97	1.14	1.21	1.26	1.13	1.34

Table IV
AVERAGE AND STD DEV OF ERRORS BY TOOL, TASK, AND DATASET

- *"Is there a correlation between the length of the runway (which could indicate time to prepare for takeoff), incident rate, and size of the aircraft?"* (DataFramer) The dataset does not contain information about runway length, as the participant has assumed.
- *"For each sky cover type, what is the ratio birds seen over the number of birds struck?"* (DataFramer) These values are represented in the dataset as ranges, which could not correctly be used to compute this ratio. Thus, the participant has made an incorrect assumption about the format of the data.

• Assuming the data means something that it does not mean i.e., reasoning incorrectly about the data semantics:

- *"In the TaxiIn column, is there more information regarding the taxi service?"* (spreadsheet) The participant has assumed that the column "TaxiIn" refers to transit via taxi cab, rather than to the movement of the aircraft on the ground in route to the airport gate.

2) Logic errors in questions, hypotheses, and plans: We found several cases.

• Incorrect interpretation of a result of a hypothetical analysis based on a misinterpretation of the results of the proposed operations that would be applied to the data:

- *"An interesting question that this data set can help us answer is what is the average height at which various bird species fly."* (DataFramer) This dataset does not provide information about where birds are flying, but rather, about where birds are struck by planes. This would not allow estimation of the average height at which various bird species fly.

• Logically nonsensical, in some cases reflecting incorrect use of vocabulary or technical terms:

- *"Challenge: The influence of DepDelay on DepTime may be different for each airport, for example, related to its being large or small."* (spreadsheet) Departure delay (DepDelay) is the result of subtracting the scheduled departure time from the actual departure time (DepTime). DepDelay therefore deterministically influences DepTime, by definition. This could not vary by airport.
- *"There are possible issues with the dataset as some numbers do not match, like why there are less canceled flights than there are flights with departed time?"* (DataFramer) This statement proposes there is some issue with there being fewer canceled flights than flights that departed. It is possible this was a typographical error and meant to point out that there are fewer canceled flights than flights without a departure time.

• Underspecified to the point of reflecting an incompletely considered analysis:

- *"Which airports suffer from most delays?"* (DataFramer) This statement is underspecified because "most delays" could mean most delayed flights, highest average delay per flight, highest cumulative delay, or other.
- *"I could also imagine answering regulatory questions based on the dataset."* (DataFramer) This statement again reflects a vagueness or uncertainty about the exact course of action or the analysis steps that would be involved, which is likely to cause the participant to struggle when they attempt to put plan into practice.

• Inappropriately generalizing from a sample size that is too small even though they recognize the size of the sample:

- *". . . it would be interesting to see what the most fatal part of a flight [is] (which could lead to more effective safety measures at those parts)."* (DataFramer) This dataset only contains two instances where fatalities resulted. However, we cannot be sure that the participant noticed this fact. Thus, to err on the side of conservatism, we mark it as a logic error, rather than a data error.

3) Data errors in observations: We found one type. For example:

"There are a total of 2329 flights that were delayed by one minute or more." (spreadsheet) There were actually 2324 flights that were delayed by one minute or more. Thus, the participant has made a calculation error.

4) Logic errors in observations: We found one type. For example:

"I attempted to look into which types of wildlife are most at risk for being hit...gulls seem to be at risk." (DataFramer) It is the case that in many of the strike incidents, gulls were involved. However, this does not necessarily mean that gulls have a particularly high chance of being struck by a plane. Rather, it could be that there are a very large number of gulls relative to other bird populations. This would contribute to the involvement of gulls in strikes while still being consistent with gulls having an overall low likelihood of being struck.

D. Other errors

Apart from the error types we just described, we observed a number of other errors. We opted to exclude these from explicit evaluation because we found it infeasible to objectively identify these errors in all cases.

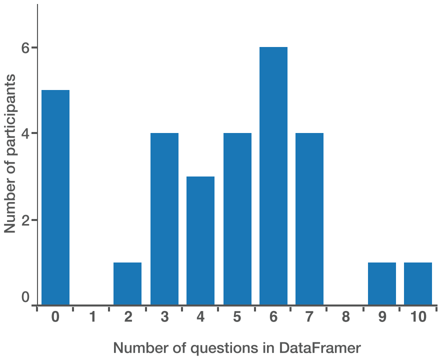


Figure 7. All but five participants used the question functionality in DataFramer, with the majority of participants posing five or more questions.

1) Base rate errors: One very common type of error happened when a participant expressed a question that they expected to be answerable with the data, without considering that a ratio of events—including data that was not available to them—would be needed to properly address the posed question. We observed this type of error more commonly in the wildlife strike dataset, especially tied to the notion of risk. These were most prevalent in the spreadsheet condition, possibly because spreadsheet formats can make it difficult to get a full grasp of the available data.

2) Generalization errors: We also observed generalization errors. These arise when someone is inappropriately generalizing from the dataset at hand to the larger context from which the data was gathered. This error was prevalent in the on-time departures dataset, which only contained one week of data from January 2010. A large number of memos written about this dataset described questions and analyses that would not be valid given the limited time frame. To be conservative, we only count it as a data error if they specifically made a comment that revealed that they did not recognize the limited extent of the dataset. We count it as a logic error if they both explicitly acknowledge the limited extent of a dataset, and explicitly generalize beyond that limited sample inappropriately. Otherwise we did not count an error. In most cases, whether or not a participant may have been inappropriately generalizing came down to subtle interpretations of language.

E. Tool usage

How do we know that the improvements we observed when participants used DataFramer came as a result of actually using the features of the tool we designed to address the problems described in Section III? For this we have to look at how extensively participants used the various aspects of DataFramer: (1) looking at charts, (2) composing questions, (3) associating questions with specific columns of the data that would be needed to answer them, and (4) reflecting upon question quality and marking them to keep or reject. We present usage data on the last three.

We found that all but five participants used the question functionality in DataFramer, with the majority of participants posing five or more questions (Figure 7). All but six participants associated specific columns with questions in DataFramer, with the majority of participants doing this for four or more questions. Almost half of participants (14) marked questions to keep or reject when using DataFramer. Most of these were marking questions as keep. Seven distinct participants each rejected exactly one of their questions.

In all but two cases, participants left the rejected questions out of their memos. Four of the rejected questions were unanswerable because the datasets do not contain the data required to answer the question. In one case, the participant even remarked in their memo that they did not have enough data to answer the question. Two rejected questions were easily answerable from the README we provided with the data or from quickly skimming the data. Perhaps participants rejected these for being relatively uninteresting or because they answered it already. This points to the need to reconsider the organizational categories provided by DataFramer; having a place to mark answered questions along with the answer could be useful. However, note one participant left one such question in their memo rather than leaving it out. Lastly, one rejected question appeared to be valid and answerable by the data, so it is unclear why the participant rejected it.

1) Tool feedback: Some participants provided unsolicited feedback about the tools in their memo, though the study protocol did not instruct them to do this.

DataFramer feedback: Five participants commented that they appreciated the automatic overviews. The main drawbacks named were:

- variables cannot be directly compared to one another (four participants),
- viewing all charts at once on the page is overwhelming (one participant), and
- small improvements are needed in label visibility, scrolling, etc. (four participants).

Future versions of the tool should address these problems.

Spreadsheet feedback: Three participants noted their frustration at trying to create histograms for every column in a spreadsheet. This supports the idea that reducing the tedium of generating such overviews is a useful feature. Three also commented that when using the spreadsheet, there were operations they wanted to carry out, such as seeing all possible values for a column, but were unable to identify the low-level steps needed to do so.

VI. Implications

Our observations imply a number of interesting takeaways, both for the design of tools for exploratory data analysis and the design of future studies. We discuss these in terms of the four challenges identified in Section I.

Lack of support for improvised workflows: Our observations indicate that encouraging people to explicitly pose questions may have encouraged them to ask slightly more questions than they would otherwise. However, we speculate that more differences would have been observed if the study protocol had not explicitly instructed participants to compose questions in their memos, because people do not usually think in terms of questions with other analysis tools. By doing that for all tools, we were able to see a common pattern of errors that did not seem to be very dependent on the tool.

Unnecessary tedium due to repetition: Five participants commented that the automatically generated overviews in DataFramer were helpful, and three that it was frustrating to create such overviews in a spreadsheet. These overviews also seemed to result in a reduction in the number of erroneous assumptions. On the other hand, too many overviews could be overwhelming to some users.

Tension between unfamiliarity and specification: When using a spreadsheet or Tableau, some participants were confused by the complexity of the provided functionality, and weren't able to able to figure out how to perform desired tasks in a timely manner. When using DataFramer, some participants were frustrated at the lack of support for certain tasks, like comparing variables. One end of the spectrum represents limiting functionality so that users stay focused. The other end represents enabling users to do nearly anything, at the risk getting sidetracked when trying to manage the complexity. Perhaps some combination of these approaches —one whose interface optionally exposes functionality and context about the dataset in a logical sequence, starting with questions and basic overviews, —could represent promising middle ground. This could temper the temptation to dive into a dataset and apply complicated analyses before obtaining important context about that dataset.

Hidden erroneous assumptions: Our observations revealed error types we did not anticipate in advance. In part for this reason, our tool does not address all types with equal effectiveness. In particular, logic errors were not addressed by DataFramer. This suggests it is important to make more information about dataset context and semantics available. Based on an intuition that this was important, we designed one early version of DataFramer to have an annotation feature for users to leave notes about individual columns. We removed this feature because, in usability tests, we observed it was almost never used. However, the logic errors suggest that some better variation of this feature, perhaps with automatically populated metadata or content information, could be helpful.

Overall, while we cannot say without a doubt that DataFramer represents an improvement in tools for exploratory data analysis, this research has identified important problems and resulted in interesting observations that can be used to improve the design of future tools and studies. We uncovered a rich set of patterns in the errors produced using both tools that instruct the design of the next round of features. Overall, the approach of examining data and stating questions before exploring the data seems promising.

References

[1] L. V. Jones, Ed., *The collected works of John W. Tukey: philosophy and principles of data analysis*. Chapman and Hall, 1986.

[2] B. Bederson, "Interfaces for staying in the flow," 2004.

[3] D. Hoaglin, F. Mosteller, and J. Tukey, *Understanding Robust and Exploratory Data Analysis*. John Wiley and Sons, Inc., 1983.

[4] L. Grammel *et al.*, "A survey of visualization construction user interfaces," *Eurographics Conference on Visualization (EuroVis)*, 2013.

[5] "Tableau Software," www.tableausoftware.com.

[6] C. Stolte, D. Tang, and P. Hanrahan, "Polaris: A system for query, analysis, and visualization of multidimensional relational databases," *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2002.

[7] J. Mackinlay, "Automating the design of graphical presentations of relational information," *ACM Transactions on Graphics (TOG)*, vol. 5, no. 2, pp. 110–141, 1986.

[8] S. Kandel *et al.*, "Wrangler: Interactive visual specification of data transformation scripts," in *Human Factors in Computing Systems (CHI)*, 2011.

[9] ——, "Profiler: Integrated statistical analysis and visualization for data quality assessment," in *Advanced Visual Interfaces (AVI)*, 2012.

[10] A. Perer and B. Shneiderman, "Systematic yet flexible discovery: guiding domain experts through exploratory data analysis," in *Conference on Intelligent User Interfaces (IUI)*, 2008.

[11] S. Few, "Exploratory vistas: Ways to become acquainted with a dataset for the first time," in *Visual Intelligence Business Letter*, July-Sept 2011.

[12] A. Inselberg, "Multidimensional detective," in *IEEE Symposium on Information Visualization*, 1997.

[13] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations." in *IEEE Symposium on Visual Languages*, 1996.

[14] D. Kurlander, "Reducing repetition in graphical editing," *HCI International Conference*, 1993.

[15] Y. Sun, "Articulate: Creating meaningful visualizations from natural language," Ph.D. dissertation, University of Illinois at Chicago, 2003.

[16] M. Voigt *et al.*, "Context-aware recommendation of

visualization components," *Conference on Information, Proces, and Knowledge Management (eKNOW)*, 2012.

[17] ——, "Using expert and empirical knowledge for context-aware recommendation of visualization components," *International Journal on Advances in Life Sciences*, 2013.

[18] A. Key *et al.*, "Vizdeck: self-organizing dashboards for visual analytics," in *Management of Data (SIGMOD)*, 2012.

[19] S. Roth *et al.*, "Visage: A user interface environment for exploring information," *IEEE Symposium on Information Visualization (InfoViz)*, 1996.

[20] ——, "Interactive graphic design using automatic presentation knowledge," in *Human factors in computing systems (CHI)*, 1994.

[21] S. Casner, "Task-analytic approach to the automated design of graphic presentations," *ACM Transactions on Graphics (TOG)*, 1991.

[22] A. Bernstein *et al.*, "Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification," *Knowledge and Data Engineering*, 2005.

[23] R. St. Amant *et al.*, "Intelligent support for exploratory data analysis," *Journal of Computational and Graphical Statistics*, 1998.

[24] M. Schiff, *Designing graphic presentations from first principles*. University of California, Berkeley, 1998.

[25] Z. Wen and M. Zhou, "An optimization-based approach to dynamic data transformation for smart visualization," *International Conference on Intelligent User Interfaces (IUI)*, 2007.

[26] M. Derthick and S. Roth, "Example-based generation of custom data analysis appliances," *International Conference on Intelligent User Interfaces (IUI)*, 2001.

[27] D. Gotz and M. Zhou, "Characterizing users' visual analytic activity for insight provenance," *Conference on Visual Analytics Science and Technology (VAST)*, 2009.

[28] D. Gotz *et al.*, "Harvest: Situational visual analytics for the masses," *Workshop on Intelligent Visual Interfaces for Text Analysis (IVITA)*, 2010.

[29] J. Lu *et al.*, "Analytic trails: Supporting provenance, collaboration, and reuse for visual data analysis by business users," *IFIP TC13 International Conference on Human-Computer Interaction (INTERACT)*, 2011.

[30] J. Heer *et al.*, "Graphical histories for visualization: Supporting analysis, communication, and evaluation," *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2008.

[31] L. Bavoil *et al.*, "Vistrails: Enabling interactive multiple-view visualizations," *IEEE Conference on Visualization (VIS)*, 2005.

[32] M. Kreuseler, T. Nocke, and H. Schumann, "A history mechanism for visual data mining," *IEEE Symposium on Information Visualization (InfoViz)*, 2004.

[33] C. Dunne *et al.*, "Graphtrail: analyzing large multivariate, heterogeneous networks while supporting exploration history," *ACM Conference on Human Factors in Computing Systems (CHI)*, 2012.

[34] D. Kurlander and S. Feiner, "A history-based macro by example system," *Symposium on User Interface Software and Technology (UIST)*, 1992.

[35] H. Yang, Y. Li, and M. Zhou, "Understand users' comprehension and preferences for composing information visualizations," *Transactions on Computer-Human Interaction (TOCHI)*, 2014.

[36] D. Toker *et al.*, "Towards adaptive information visualization: On the influence of user characteristics," *Conference on User Modeling, Adaptation, and Personalization (UMAP)*, 2012.

[37] A. Bigelow *et al.*, "Reflections on how designers design with data," *Advanced Visual Interfaces (AVI)*, 2014.

[38] A. Perer and B. Shneiderman, "Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis," in *Conference on Human Factors in Computing Systems (CHI)*, 2008.

[39] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical Studies in Information Visualization: Seven Scenarios," in *Transactions on Visualization and Computer Graphics (TVCG)*, 2012, pp. 1520–1536.

[40] Y. A. Kang *et al.*, "Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study," in *Visual Analytics Science & Technology (VAST)*, 2011.

[41] ——, "Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study," in *Visual Analytics Science & Technology (VAST)*, 2009.

[42] Y. ah Kang and J. Stasko, "Examining the use of a visual analytics system for sensemaking tasks: Case studies with domain experts," *Transactions on Visualization and Computer Graphics (TVCG)*, 2012.

[43] C. Weaver, "Conjunctive visual forms," *Transactions on Visualization and Computer Graphics (TVCG)*, 2009.

[44] F. Lieder *et al.*, "Algorithm selection by rational metareasoning as a model of human strategy selection," *Neural Information Processing Systems (NIPS)*, 2015.

[45] T. Jankun-Kelly, K.-L. Ma, and M. Gertz, "A model for the visualization exploration process," *Transactions on Visualization and Computer Graphics (TVCG)*, 2007.

[46] H. Wickham, "My advice on what you need to do to become a data scientist..." 2015, https://gist.github.com/hadley/820f09ded347c62c2864.