Sophia Hsuan-hsin Huang
16765514
INFO 247 - FINAL
MAY 2021

LUX - NEW INTENT

PROJECT GOALS

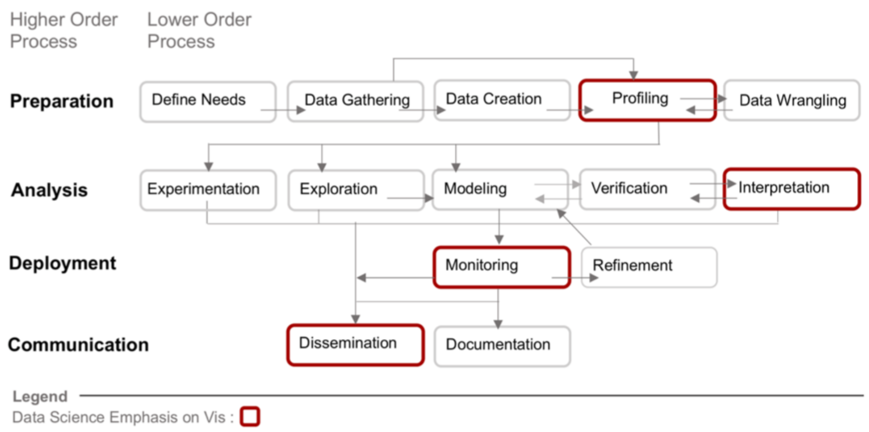
The initial iteration of Lux provided limited support for databases intended for larger datasets such as Postgres. To identify and prioritize additional SQL Executor features that would make Lux more suitable for database users, we conducted qualitative user research by leveraging remote interviews of potential Lux users. Additionally, a secondary objective of the study was to understand the workflow of Lux's target and adjacent user personas to uncover the pain points that their current visualizations tools are unable to alleviate, thereby increasing the usability and applicability of Lux to a broader audience.

To better inform the technical work of our Capstone, which aimed at fleshing out Lux’ database functionality, we wanted to have a clearer idea of what features users would want out of a data exploration tool and their feedback on our initial prototype. Thus, the goal of our user experience studies was to identify users’ pain points in their data exploration workflows and validate potential features that could be integrated with the existing version of Lux.

Lux - New Intent falls under a subset of the overarching goal of broadening the user base of Lux to those who are not traditionally recognized as data scientists yet conduct exploratory data analysis as part of their regular workflow. We wanted to explore more intuitive, interactive approaches to generate visualizations. The goal is to understand if the interactive questionnaire is a feasible, valuable method to proceed further, compared against existing Lux’s current avenues.

RELATED WORK

Visualization for Profiling vs Experimentation and Exploration[1]



Crisan, Fiore-Gartland, and Tory explains how data visualization used in the experimentation and exploration stages of analysis is fundamentally different from the goals of the profiling stage. The motivation to create data visualization in profiling is to assess data quality and understand data content on a high-level, such as distribution, identification of missing values, and associations between attributes. By contrast, data professionals use visualizations to explain causal inferences in experimentation and "seeks to uncover new insights from data that, unlike in experimentation, were not predetermined from the outset" during exploration.[2]

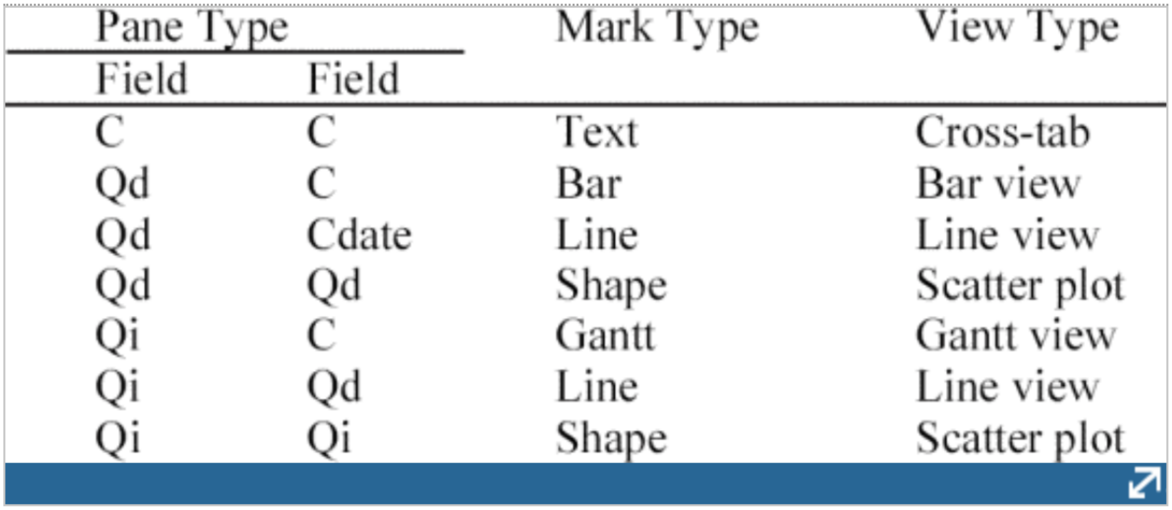
It was important for us to understand the distinctions between the different usages of data visualization throughout the lifecycle of data work on the ground. We chose to focus the project on improving the experimentation and exploration aspects. Moreover, it informed the decision to use a sequence of questions in the interactive questionnaire prototype, comparable to hypothesis formation.

Show Me: Automatic Presentation for Visual Analysis[3]

[1] https://research.tableau.com/sites/default/files/Crisan_DataScience.pdf
[2] Same as above, page 3.
[3] https://ieeexplore.ieee.org/abstract/document/4376133?casa_token=efSD_jLdhf8AAAAA:yjiFlCeKyXnmhcPqLsBSJYOaDTu_iSMCylpmKGCUeSGmwS2uFOC1guoGt8lkGLT37jkibV6u980

Mackinlay, Hanrahan, and Stolte developed automatic presentation functionality and previews of small multiple views using heuristics - affinity, multiple fields, and add to sheet. Specifically, we drew inspiration from Tableau's data model on its approach to data properties, i.e. data types, data roles and data interpretation, and automatic marks of classifications, i.e. categorical or quantitative, to suggest the view type or visualization. This helped shape the specific data requirements to gather and the order of the requirements in the interactive questionnaire so a mapping of visualization could be made.

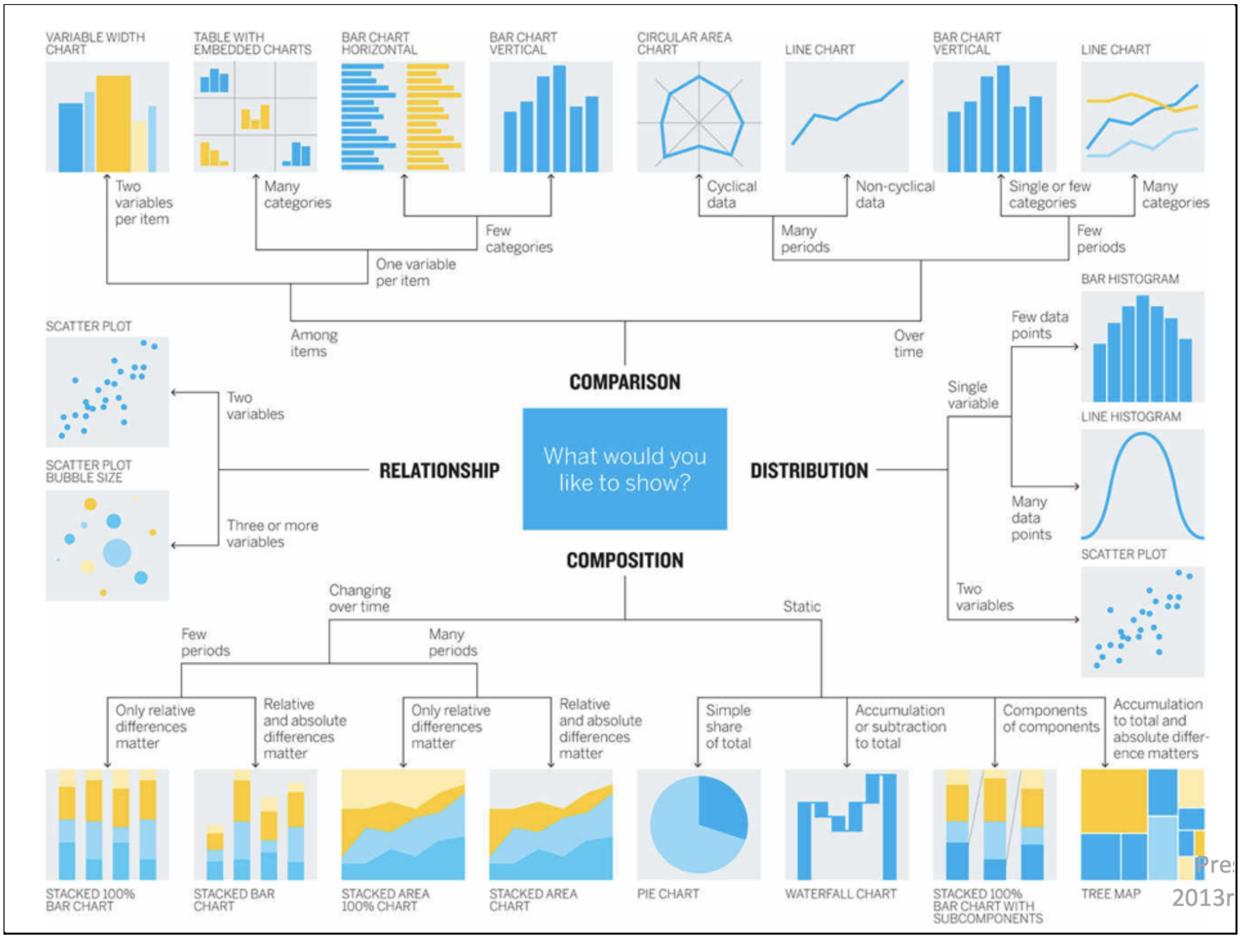
TABLE 1 Automatic marks rules



Pane Type		Mark Type	View Type
Field	Field		
C	C	Text	Cross-tab
Qd	C	Bar	Bar view
Qd	Cdate	Line	Line view
Qd	Qd	Shape	Scatter plot
Qi	C	Gantt	Gantt view
Qi	Qd	Line	Line view
Qi	Qi	Shape	Scatter plot

Andrew V. Abela, Advanced Presentations by Design, 2013 Redrawn by Berinato. Good Charts

Figure 1: Charts of visualizations by relationships



The first prototype was born out of inspiration from coursework from Information Visualization and Presentation (INFO 247 Spring 2021) taught by Professor Marti Hearst and the diagram from Andrew V. Abela redrawn by Berinato. We started to converge around the idea of how we might translate user intent into a series of specific data gathering requirements to map it to a set of recommended visualizations. After consulting with Professor Hearst, we agreed that the prototype could be limited to the scope of Distributions (center right section of the diagram) because a) the visualizations under distributions are all visualizations supported by Lux's current capabilities and b) any interface interaction introduced anew could be compared against Lux's current Distribution tab feature.

Multiple-Comparison Data Modeling [4]

Building on the ways that p-hacking may erode research validity, Gelman and Loken propose using multilevel modeling and thorough analysis of relevant comparisons of the dataset from the outset of the data exploration process to resolve these multiple-comparison issues.[5] The authors state, "A starting point would be to analyze all relevant comparisons, not just focusing on whatever happens to be statistically significant."[6] For the context of EDA on largely observational studies in political science or economics, having a rigorous approach to analyzing data with a clear statistical framework of multilevel modeling alongside hypothesis formulation would mitigate errors arising from "insufficient modeling of the relationship between the corresponding parameters of the model."[7]

In many ways, while we understand one of Lux's primary goals is to automate EDA quickly and present visualizations instantly to the user, there is potential in having a more methodical interaction method that can help researchers reduce missteps in procedure prematurely. The more deliberate approach behind the iterative phases of the following intent prototypes attempts to erect guards against aforementioned examples of misuse. Additionally, adding more structure to the EDA process that Lux automates invisibly could complement Lux by giving the user more intuitive, visible control and precision over their own goal-driven actions. As summarized and defined by Dimara and Perin, "good interaction minimizes error and distance to user goal, and provides rapid and stable convergence to the target state."[8]

VISUALIZATION DESCRIPTION

[4] http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

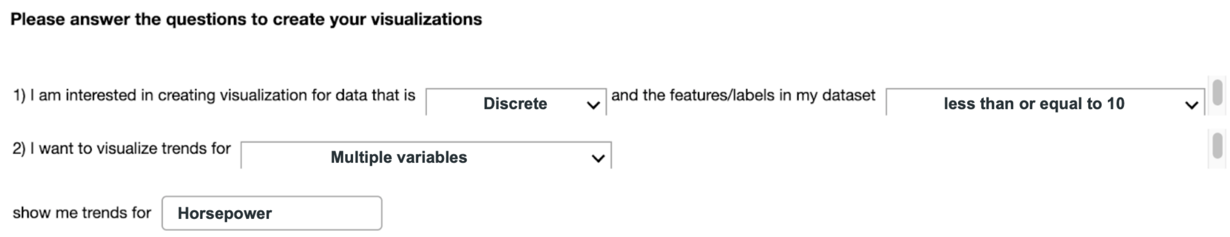
[5] Same as above

[6] Same as above, page 14

[7] Andrew Gelman, Jennifer Hill & Masanao Yajima (2012) Why We (Usually) Don't Have to Worry About Multiple Comparisons, Journal of Research on Educational Effectiveness, 5:2, 189-211, DOI: 10.1080/19345747.2011.618213, page 190

[8] Evanthia Dimara, Charles Perin. What is Interaction for Data Visualization?. IEEE Transactions on Visualization and Computer Graphics, Institute of Electrical and Electronics Engineers, 2020, 26 (1), pp.119 - 129. ff10.1109/TVCG.2019.2934283ff. ffhal-02197062

Figure 2: Version 1 of Interactive Questionnaire



Please answer the questions to create your visualizations

1) I am interested in creating visualization for data that is **Discrete** and the features/labels in my dataset **less than or equal to 10**

2) I want to visualize trends for **Multiple variables**

show me trends for **Horsepower**

First, the interactive questionnaire is embedded in Lux's current system of design and inside the jupyter notebook so it feels natural for the user to interact. Moreover, the questionnaire is added as an additional tab next to the leftmost one to complement the existing approach to visualization generation.

We created two versions of a similar approach, and gathered feedback.

Version 1

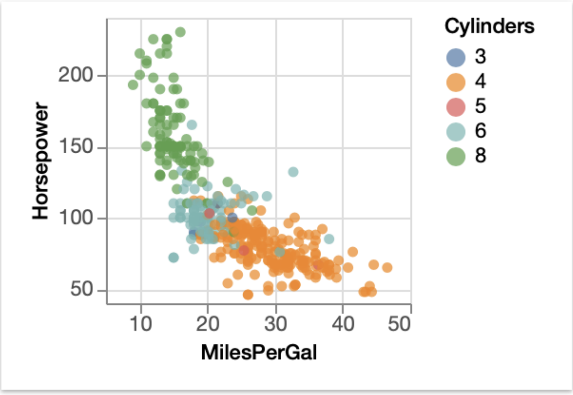
We began with asking the user to fill out some questions that will result in outputs of visualizations. We structured the questions with increasing specificity of detail on data: starting from data type (discrete or continuous) and total size of attributes, then determining the attributes for the visualization, then specifying the exact attributes of interest.

Table 1: Table of Details of Visualization's User Interaction

Content of Selections	**User Interaction**	**Description**
discrete vs continuous	Dropdown menu	Gather data type
less than or equal to 10 vs more than 10	Dropdown menu	Gather total size of dataset for performance goals
One vs two vs multiple variables	Dropdown menu	Gather data role to determine view type
Horsepower	Text field box	Limit to specific attributes of interest using panda dataframe

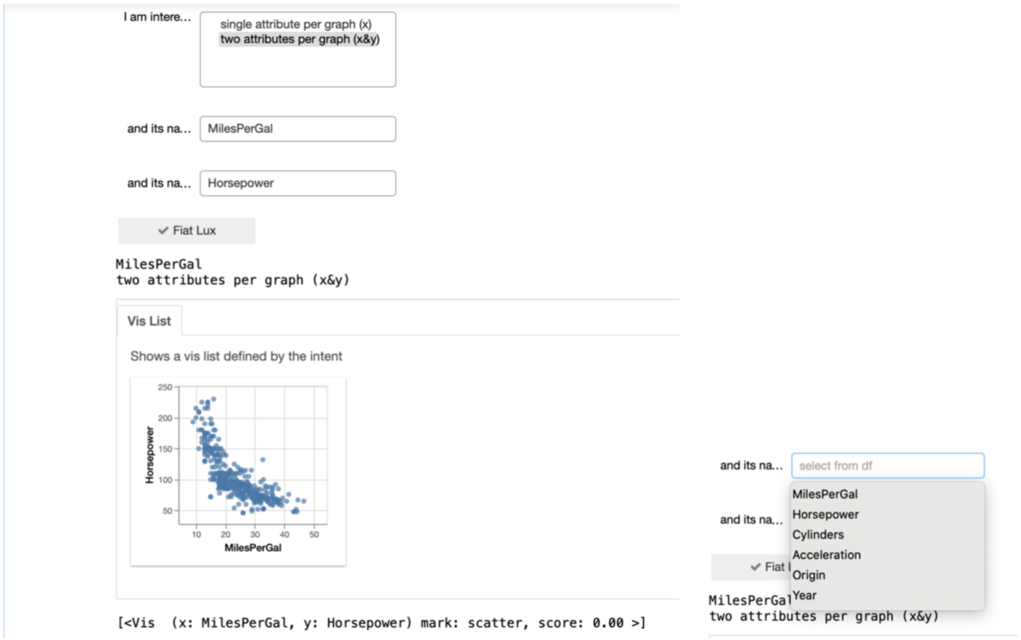
We attempted to word the questions in a more conversational tone yet for the iteration we tested for usability testing, we realized the tone sounds more command-like and intentional. Once the user finishes form filling, they click on the Fiat Lux button to generate a visualization or sets of visualizations below. The following visualization is an example.

Figure 3: Generated Visualization from Interactive Questionnaire



Version 2

Figure 4: Version 2 Interactive Questionnaire



I am intere... single attribute per graph (x) / two attributes per graph (x&y)

and its na... MilesPerGal

and its na... Horsepower

```
MilesPerGal
two attributes per graph (x&y)
```

Shows a vis list defined by the intent

```
[<Vis  (x: MilesPerGal, y: Horsepower) mark: scatter, score: 0.00 >]
```

and its na... select from df

and its na...

```
MilesPerGal
two attributes per graph (x&y)
```

What differentiated this version 2 from 1 was both its vertical format and surfacing the user inputs all at once, with the ability for automatic form-field population beyond after the first few answers. We thought users would prefer to have all tokenized answers and attributes from panda dataframe visible to better inform their selections. We also attempted to print the interestingness score calculations and the mark types, channels, labels, and encodings that make up a visualization to improve transparency of the inner workings of Lux.

APPROACH

Data, Tools, and Steps

Understanding we needed to create the interactive questionnaire inside the jupyter notebook, we relied on the same IP display used by Lux--ipywidget. Then we developed a questionnaire in a jupyter notebook with Python using the generic cars dataset as a proof of concept.

Through the early stages of the prototype, we also relied on a combination of ipywidget and figma to design the interactive interface for iterative rounds of user testing.

USER TESTING AND RESULTS

Overview

During the course of the Lux UX study, we interviewed six participants. The coordinators leveraged three channels: UC Berkeley Information School Slack groups, alumni, and professional networks, to recruit potential interviewees. We sought individuals who had some experience with data science, worked with databases in their day-to-day or were interested in having an easy-to-use visualization solution. For this study, there was no screener to validate participants. However, to garner interest from target participants, a blurb with potential characteristics of Lux's intended audience, its value proposition, and a GIF of Lux was broadcasted across the aforementioned slack groups. In addition, before the interviews, participants were given a brief overview of the Lux to familiarize them with its capabilities and features. Recruited participants had a mix of roles, but each participant demonstrated advanced coding proficiency.

Method & Test Setup

The project incorporated a design thinking approach to identify and address user pain points through the following steps:

1. Empathize: Through exploratory interviews of Lux users and similar personas, we gathered intimate knowledge about gaps in the industry, the goals of the users, and their reservations against Lux.

2. Define & Ideate: The insights and findings from interviews informed the ideation process where the team gathered to scope the problem and brainstorm solutions.

3. Prototype & Test: Using Figma and ipywidgets initial prototypes were designed to gather feedback from the users.

GIven the interviewees had different prior level of familiarity with Lux and coding proficiency during the usability testing, we focused on the end-to-end scenario as broadly defined EDA operations and ask the same questions for 3 iterations:

1) Perform custom and investigative data query around a key metric or dataset with the goal to generate data visualization to answer related questions
2) Load dataset into Lux
3) Select a few attributes or variables of interest among all attributes or variables
4) Describe relationship of interest for dataset
5) Select a visualization or a set of visualization among many presented
6) Export or download the selected visualization or set of visualizations

User Persona

Data Scientist:

A Data Scientist creates machine learning models using data gathered from different product teams.

Goals:

1. Interprets model outputs to inform the product team.
2. Share insights with stakeholders.

Common Frustrations:

1. Exploratory Data Analysis takes ~ 10% of work time.
2. Libraries like Pandas don't offer ease of use and granularity at the same time.

Business Analyst:

A Business Analyst identifies insights from the data to help stakeholders make business decisions.

Goals:

1. Identify data trends to validate & inform business decisions.
2. Share insights with stakeholders.

Frustrations:

1. Creating situation specific presentable visualizations.
2. Limited integration across different tools.

Image source[9]

Hypothesis of User Testing

1. There is a spectrum of intent specificity, ranging from abstract to specific, that is currently unmet by Lux's current technical capabilities

[9] Image Source: https://icons8.com/illustrations/author/5eb2a7bd01d0360019f124e7 by Olha Khomich from icons8.com

2. There is room for improvement regarding Lux's current interface interactions. These can be more natural language based and less code dependent

3. There is a need for greater control over the sets of recommended visualizations through customized interestingness metrics or user controls

Design Heuristics

From the hypothesis, we formulated a few ideas and designed the prototypes around these design heuristics/principles:

- Seamless integration with Lux
- Language-based or visual-based interface interaction
- Clarity of user progress
- No coding required
- Greater user control and accessibility

Result Analysis

KEY THEMES WITH USER INTERVIEW QUOTES

1. SUMMARY STATISTICS, BOXPLOT

● "I am a fan of boxplots. It's the simplest way to quickly read the data's summary statistics visually." (P1)

2. CONVENTIONAL, REPETITIVE OPERATIONS OF EDA, I.E. MISSING DATA

○ "Restructuring the data in the way I want is the biggest frustration I have and sometimes even after uploading it into data visualization softwares, it still doesn't create the graph that I want.” (P6)

3. TRANSPARENCY AND USER CONTROL OF ALGORITHMIC RANKING/OUTPUT

○ “It was not clear the visualizations were ranked. So maybe, even like, when those plots are being produced, you know, where it shows the relationship between two quantitative attributes, just literally saying, like, ranked by correlation, or just literally telling me like, what, what the ranking method is, will help me understand.” (P5)

○ “Only show me the visualizations that fit within a certain range or specified threshold of the algorithm, which reduces the amount of things that I have to look at.” (P6)

4. NAVIGATION

○ “Give me as much graph as the window view fits. Don’t make me scroll right and left to look for graphs if I don’t have to.” (P6)

○ “You can hide some graphs but still give me the option to see more or all graphs if I want to.” (P1)

5. AUTO-DETECTION AND ANTICIPATION OF COMMON DATA PROBLEMS (TEMPORAL DATA)

○ “Suppose if one in a million numbers is a string and pandas convert it to string. It would be nice if it would alert you like, hey, one line out of your, like, million is messed up. And these strings really are numbers.” (P5)

○ “For instance, how do you extract the zip code out of [a dataset] or how to determine an address from a blob of text. Trifecta gives you a quick and easy way to understand that this is the city zip code." (P5)

Overall Impression

What emerged was a set of target users who are primarily embedded in the technology team and shared similar patterns of data-driven workflows without the formal job title as "data scientist."

As we learned about their day-to-day tasks and identified their main responsibilities, it became clear:

1. Data analysis was no longer a job exclusive to the explicit role of data scientists.
2. Basic data queries are often conducted within each functional team without assistance from the company's core data science department.
3. For companies of 5,000 and above, incorporating data-driven methods in decision-making or product development cycles is today's cultural norm.

For usability results on the new intent prototype, all users expressed excitement about the questionnaire approach to EDA. Among the highlights, its seamless integration into Lux's existing design systems, intuitive interactions, and series of natural language questions mapped to paths of visualization generation offered another complement to strengthen Lux's adoption usage beyond traditional circles of data scientists. More importantly, users' expression of intent have become much more defined, from specifying the variables of interest to formulating sound hypothesis structure to orchestrating multiple-comparison modeling even before they begin EDA, afforded by the sequence of guided prompts. Users spoke to the structure and sequence of questions that informed their thought trajectory in the questionnaire to see its promising potential. While we did not find out how the new questionnaire approach fared over comparison with the current auto-generated sets of visualizations underneath the Distribution tab, we learned major areas of improvements that could be incorporated into future enhancements:

1) Increase variety of types of visualizations that are critical and common to EDA i.e. Boxplot or Violin plot.

2) Simpler handling method or streamlining of data preprocessing and wrangling helps users get to making visualizations faster. Widespread frustration with unavoidable, conventional, and repetitive operations for multiple variables within the same dataset is a huge pain point for EDA. To double-down on Lux's original value proposition of fast and simple EDA, it may be valuable to look into conventional operations of EDA and display summary statistics alongside visualizations.

3) Greater transparency and user control on algorithmic rank with code statements or interactive refining components. Some proposed solutions were more details for calculation of the interestingness function and its equation, or slider of the range of numeric values of the function.

4) Better navigation and larger display window to see resulting outcomes of visualizations in a single view, and with the option to see more, e.g. an accordion signifier that opens up to more visualizations when they don't fit in a single view.

5) Automatic detection of data types and communication of potentially problematic subsets of data upfront.

LINK TO CODE

GitHub

https://github.com/lux-org/lux

Attached notebook

DIVISION OF LABOR

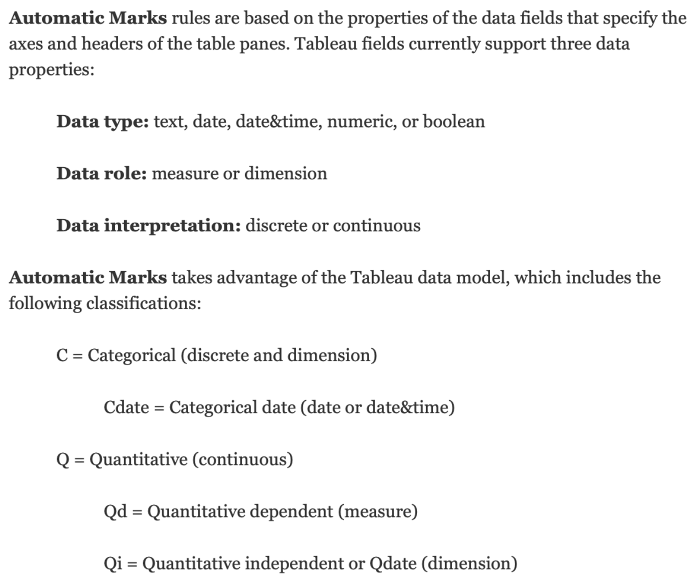
I have contributed to any aspects that touch on user interviews, research, and design.

- User interviews
- Design prototypes
- Usability write-up
- Final write-up

APPENDIX

Show Me: Automatic Presentation for Visual Analysis[10]

Figure 2: Tableau Data Model



Automatic Marks rules are based on the properties of the data fields that specify the axes and headers of the table panes. Tableau fields currently support three data properties:

Data type: text, date, date&time, numeric, or boolean

Data role: measure or dimension

Data interpretation: discrete or continuous

Automatic Marks takes advantage of the Tableau data model, which includes the following classifications:

C = Categorical (discrete and dimension)

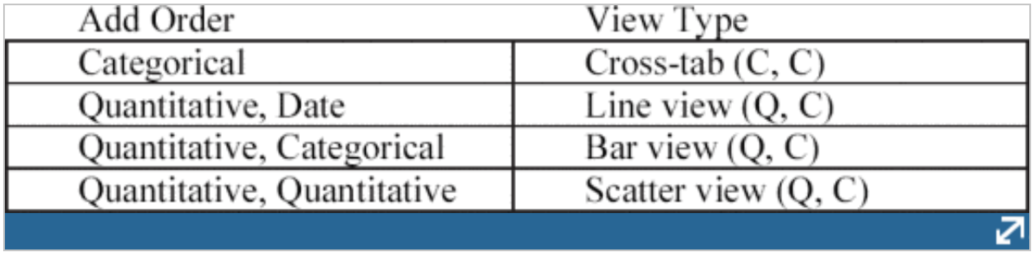
Cdate = Categorical date (date or date&time)

Q = Quantitative (continuous)

Qd = Quantitative dependent (measure)

Qi = Quantitative independent or Qdate (dimension)

Figure 3: Add to Sheet Heuristics



Add Order	View Type
Categorical	Cross-tab (C, C)
Quantitative, Date	Line view (Q, C)
Quantitative, Categorical	Bar view (Q, C)
Quantitative, Quantitative	Scatter view (Q, C)

[10] https://ieeexplore.ieee.org/abstract/document/4376133?casa_token=efSD_jLdhf8AAAAA:yjiFlCeKyXnmhcPqLsBSJYOaDTu_iSMCylpmKGCUeSGmwS2uFOC1guoGt8IkGLT37jkibV6u980