Exploring Data Exploration: Pilot Studies to Characterize Exploratory Data Analysis Activity in Tableau

Sara Alspaugh
@salspaugh
AMPLab, EECS Department, UC Berkeley

in collaboration with: Marti Hearst, Randy Katz, and the students of INFO 290-09, Fall 2014
Outline

• Overview
  – Research seminar pilot studies
  – Goals
• Part 1: Analyzing data
  – Tableau log data
  – Transcript data
• Part 2: Building applications
  – DataFramer
  – ShowMe++
Outline

• **Overview**
  – Research seminar pilot studies
  – Goals

• **Part 1: Analyzing data**
  – Tableau log data
  – Transcript data

• **Part 2: Building applications**
  – DataFramer
  – ShowMe++
August 2014:
Announced Masters-level research seminar
to study exploratory data analysis
in visual analytics tools

Background
Course focus

- 3 Tableau assignments \times 5 data sets
- 2 projects:
  - analyze EDA data, or
  - create tool for EDA
Data collection

- assignments
- presentations
- student
- transcript

records steps

works on assignments

generates events

Log:

```
```
Transcripts format

EDA ASSIGNMENT 1 - FIRST FORAY
DATASET: OAKLAND CITY BUDGET
HYPOTHESIS: THE CITY SPENDS MOST OF ITS PUBLIC SAFETY BUDGET (POLICE + FIRE) ON SALARIES.

EXPLORATION LOG

<table>
<thead>
<tr>
<th>GOAL/QUESTION</th>
<th>ACTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. See spending by department</td>
<td>Drag &quot;Amount&quot; to Rows, Drag &quot;Department&quot; to Columns</td>
</tr>
<tr>
<td>2. What do all these fields mean?</td>
<td>Swap &lt;i&gt;df&lt;/i&gt;, New Sheet (2), Drag &quot;Department&quot; to Rows, Drag &quot;Act Description&quot; to Rows, Add &quot;Account Category&quot;</td>
</tr>
<tr>
<td>3. Now I see &quot;Salaries in some values&quot;</td>
<td>Now I see which are &quot;Expense vs Revenue&quot;</td>
</tr>
<tr>
<td>4. Return to Sheet 1, Drag &quot;Account Type&quot; to Filters, Include only &quot;Expense&quot;</td>
<td></td>
</tr>
</tbody>
</table>

1. The biggest department is "Non-Departmental" - What’s that about?

2. The next biggest is "Fire" - Add new Sheet (3), Drag "Department" to Filters, Include "Non-Departmental"
Data collection

Transcript

Student

Records steps

Assignments

Works on assignments

Presentations

Log

Tableau

Generates events

8
Using logs to understand behavior

• **Pros:**
  – Usually widely available, cheap to collect
  – Capture information about system state

• **Cons:**
  – Usually for debugging and performance monitoring, not analyzing user behavior
  – Dirtiness, noise, and missing information

Tableau log format

Events are JSON objects.

```json
{
    "action": "tabui:drop-ui",
    "parameters": {
        "drag-description": "",
        "drag-source": "drag-drop-schema",
        "drop-target": "drag-drop-viz",
        "field-encodings": [
            {
                "encoding-type": "invalid-encoding",
                "fn": "[On_Time_On_Time_Performance_2001_9 Extract].[sum:Number of Records:qk]"
            }
        ],
        "is-copy": "false",
        "is-dead-drop": "false",
        "is-right-drag": "false",
        "shelf-drag-source-position": {
            "is-override": false
        },
        "shelf-drop-context": "none",
        "shelf-drop-target-position": {
            "is-overlay": true,
            "is-override": false,
            "shelf-pos-index": 0,
            "shelf-type": "rows-shelf"
        },
        "target-sheet": "Sheet 1"
    },
    "timestamp": 1409793051.437
}
```
Analysis scripts, code notebooks (IPython)
Datasets analyzed in the course with metadata
Tableau logs, assignment transcripts
Event parsing and interface context regeneration
Documentation of information missing from logs
Goals

1. Pilot studies via the research seminar
   – Identify EDA challenges
   – Assess data quality
   – Evaluate data collection approach

2. Ongoing empirical analysis
   – Identify patterns in EDA activity
   – Identify patterns in EDA reasoning
   – Relate EDA reasoning to activity

3. Applications and future work
   – Build EDA tools to address observed challenges
Outline

• Overview
  – Research seminar pilot studies
  – Goals

• Part 1: Analyzing data
  – Tableau log data
  – Transcript data

• Part 2: Building applications
  – DataFramer
  – ShowMe++
Basic properties of the data collected

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>14</td>
</tr>
<tr>
<td>Assignments</td>
<td>3</td>
</tr>
<tr>
<td>Projects</td>
<td>11</td>
</tr>
<tr>
<td>Explored datasets</td>
<td>5</td>
</tr>
<tr>
<td>Tableau logs</td>
<td>21</td>
</tr>
<tr>
<td>Events</td>
<td>4313</td>
</tr>
<tr>
<td>Event types</td>
<td>217</td>
</tr>
</tbody>
</table>
What types of events are there? How many times does each appear?

*Observations:* Steep long tail. Is this a property of the way activity is mapped to event keys or an inherent property of exploratory data analysis actions?
Takeaway: There are hundreds of events. On a closer look, we probably don’t care about all of them. We should re-tool the way events are broken down to consider the parameters.
Types of actions we care about: changing location of columns, creating columns, creating, deleting, and renaming sheets, undo, and redo.

**Observation:** Even here there is a long tail.
What order are actions performed in?

Markov diagram showing transition frequencies between events.
What actions are precede an “undo”?

**Observation:** Most common event to precede an undo is an undo!

**Next steps:** What state is being undone? What are the types of drop events undone?
Midterm Project:
Taxonomy of Visual Analytic Activity

based on:


taxonomy by Hassan Jannah
visualization by Shubham Goel
to access: shubhamgoel.me/viz/ADA/taxonomy.html
bl.ocks.org/shubhgo/80323b7f3881f874c02f
Midterm Project: Analysis of Undo

Undo Action Taxonomy

Sum of Number of Records for each Assignment broken down by Action-L1, Action-L2 and Action-L3. Color shows details about Action-L3. The marks are labelled by sum of Number of Records. The data is filtered on Participant, which leaves 11 of 11 members.
Assignment #2 Description

- **Data:** Florida 2000 presidential election county voting data
  - each row represents a county
  - contains county named, technology used, number of columns on the ballot, votes per candidate, under votes, over votes
- **Prompt:**
  - How does voting machine technology relate to voting errors and the candidate voted for?

```
> head datasets/fl-ballot-2000/fl-ballot-2000.csv
county,technology,columns,under,over,Bush,Gore,Browne,Nader,Hagelin,Buchanan,McReynolds,Phillips,Moorehead,Chote,McCarthy
Alachua,Optical,1,217,105,34124,47365,658,3226,4,20,21,0,0
Baker,Optical,1,79,46,5610,2392,17,53,0,3,73,0,3,3,0,0
Bay,Optical,1,541,141,38637,18850,171,828,5,18,248,3,18,27,0,0
Bradford,Optical,2,41,695,5414,3075,28,84,0,2,65,0,2,3,0,0
Brevard,Optical,1,277,136,115185,97318,643,4470,11,39,570,11,72,76,0,0
Broward,Votomatic,1,4946,7826,177902,387703,1217,7104,54,135,795,37,74,122,0,0
Calhoun,Optical,1,78,0,2873,2155,10,39,0,1,90,1,2,3,0,0
Charlotte,Optical,2,170,2985,35426,29645,127,1462,6,15,182,3,18,12,0,0
Citrus,Optical,1,154,54,29767,25525,194,1379,5,16,270,0,18,28,2,0
```
What columns did students focus on?
How popular was each column?
What are the column access patterns across time and student?
schema
encoding
filter
row
column

not pictured: measures shelf (appears when “Measure Values” are placed on shelf)
Where do these columns get placed?

- **Schema**
  - not pictured: measures shelf (appears when “Measure Values” are placed on shelf)

- **Row**
  - column
  - filter
  - encoding
How is placement influenced by column properties?

Not pictured: measures shelf (appears when “Measure Values” are placed on shelf)
"action": "tabui:create-calculation-ui",
"timestamp": 1409793211.666,
"parameters": {
  "use-selector": "true"
}

Missing data from logs

1. Which aggregate is changed when an aggregate is changed
2. What field is removed when a field is removed
3. What sheet a field is added to when added via right-click
4. The calculation used to produce a calculated field
5. Which sheet is copied when a sheet is copied in a certain way
6. What the current sheet is
7. What sheet is deleted when a sheet is deleted
8. What sheet is copied via right-click copy
9. What dashboard a sheet is added to
10. Which sheet show me is applied to
11. Properties of the visualizations created
12. ...
Dealing with data quality issues

1. Try to infer missing information from other information in the logs.
2. Use supplementary approach to record information that is left out from the logs.
3. Change how things are logged.
Analysis summary

• Analysis challenges:
  – events that aren’t of interest
  – missing information

• What we can learn:
  – Analysis action patterns
  – Properties of the data that users focus on
  – Problems users encounter
  – Suggestions users might be interested in
Analysis action patterns

Properties of the data users focus on

What can we learn from this data?

Problems users encounter

Suggestions users might be interested in

ShowMe++?
Lines of reasoning

- Canonical statements
  - facts
  - assumptions
  - questions/hypotheses
  - observations
  - inferences
  - conclusions

Transcripts

Encode

Visualize
What does the data mean?

Interpret the metadata.

What are “over” and “under”?
What is the relationship between “technology” and errors?

“Votomatic” had the highest count of errors.

Which “technology” had the most errors?

“over” and “under” represent types of errors.

“Votomatic” and “Optical” were associated with the highest count of total votes.
Conclude
Outline

• Overview
  – Research seminar pilot studies
  – Goals

• Part 1: Analyzing data
  – Tableau log data
  – Transcript data

• Part 2: Building applications
  – DataFramer
  – ShowMe++
EDA Challenges

- Goal uncertainty
- Complex ad hoc planned workflow management
- Declaration-implementation mismatch
Complex Ad Hoc Workflows

GraphTrail

Dunne et al. “GraphTrail: analyzing large multivariate heterogeneous networks while supporting exploration history.”

VisTrails

Callahan et al. “VisTrails: visualization meets data management.”
Knowing What But Not How

After selecting “sort descending”  

Trying to sort by bottom value

Step 25: (9/11/2014 6:05:19 PM) User left click on "Tableau - Book2 (window)" in "Tableau - Book2" 

Comment: This is not right!........ Ahhh, I need to make measure values

This was NOT what I wanted
- RCDR
- undo
- I can use this to filter by dept
- add quick filter for dept
- no, I cant
- remove
- duplicate sheet 5
Final Project: DataFramer

project by Anna Swigart and Ian MacFarland
guidelines for data exploration
Perer and Shneiderman. “Systematic yet flexible discovery: guiding domain experts through exploratory data analysis.”

Horvitz et al. “The Lumière project: Bayesian user modeling for inferring the goals and needs of software users.”

intelligent help systems

automatic analysis generation

Jock Mackinlay. “Automating the Design of Graphical Presentations of Relational Information.”

predictive interfaces

Kandel et al. “Wrangler: interactive visual specification of data transformation scripts.”

Fig. 10. The Encodes relationship in the visualization domain language. The graphical sentence is on the right. The gray lines indicate which values are encoded by the objects, and the tuple of each line is the inverse of the Encodes relationship. The bulleted list on the left defines the Encodes relationship.

\[
\begin{align*}
\text{Encodes}(o_i, a_i, \text{HorzPos}) & \Rightarrow b_i = \text{scale} \times (\text{Position}(o_i, h) + \text{offset}) \land \\
\text{Encodes} & (\text{Position}(o_i, h), r(a_i, b_i), \text{HorzPos}).
\end{align*}
\]
ShowMe++?

Current ShowMe: what visualizations are possible from the selected columns?

Goal: what visualizations are relevant for my data set?
Analysis action patterns

Properties of the data users focus on

What can we learn from this data?

Problems users encounter

Suggestions users might be interested in

ShowMe++?