Data Mining and Visualization Workshop with Matthew Lincoln

Michelle Smith Collaboratory for Visual Culture

October 29, 2014

In this workshop, PhD student Matthew Lincoln demonstrated the use of plot.ly, an online platform for sharing datasets, in creating visualizations of a data set of the National Gallery of Art's Dutch collection.

The Data Analysis Pipeline:

1. **Acquire**: The first step in mining and visualizing data is acquiring the information you want to work with, which Matt had already done for us by scraping the html-formatted metadata of the National Gallery of Art's Dutch collection from their website.
2. **Store**: Organize this data into a spreadsheet with fields for the various types of data you're interested in visualizing (ex. creation date, artist name, acquisition date, genre, size, etc.)
3. **Clean**: This is one of the most time-intensive steps in the process. Before you can create visualizations of your data, you need to make sure the fields are in identical and workable formats. (ex. All dates presented in YYYY-MM-DD format.)
4. **Mutate**: In order to ask certain questions of your data, you may need to generate entirely new variables, such as genre or the total area of a painting.
5. **Visualize**: The fun part! This particular workshop focused on manipulating a data set that Matt had already cleaned and loaded into plot.ly.
6. **Narrate**: Contextualize and describe the data visualizations you've created within the framework of other relative art historical knowledge.

Elements of a Data Table:

- **Variables**: The columns of the table; the distinct types of information known for each observation.
- **Observations**: The rows of the table (one for each object or unit of study).
- **Values**: The cells of the table (one value for each variable and each observation).

Data Types:

Different types of visualizations are best for certain types of data: Ordinal vs. Ordinal (scatterplot), categorical vs. ordinal (bar chart), ordinal or categorical distributions (histogram).

- **Categorical**: Data with a limited number of possible values (ex. Genre or Collector).
- **Ordinal**: Data comprising numerical scores along an ordered scale

We worked through two exercises during this workshop, the first explored the NGA's Dutch Collections, and the second looked at sales of British art from the Getty Provenance Index.

First Exercise: Using histograms, bubble charts, scatterplots, and box charts, we considered the distribution of the NGA's Dutch acquisitions, relative to creation date, acquisition date, and collector (Mellon, Widener, and Wheelock).

From the plot.ly home screen, click on the variables you're interested in visualizing, set your X and Y axes, and at left, select the type of chart you want to create. (When you're first using plot.ly, it may automatically select variables for you. Scroll all the way across your data set to make sure only the variables you want to show are selected, and unselect any unwanted variables.)

What if we want to see how the NGA collection represents the seventeenth-century production of Dutch art? Start by plotting creation date alone.

Plot.ly automatically organizes the histogram by decade. Once you've created your visualization, click 'Traces' at the upper left and decrease the 'bin size' to show a wider numeric range. The smaller the bin size, the more specific your visualization will be. A large bin size may mask some discrepancies in the data.

What if we want to see how the creation dates of objects in the NGA's Dutch collection relate to the collectors who acquired them? Go back to the data table on plot.ly's home page. Add color to your histogram to distinguish the acquisitions by collector! Click 'Group By' at left and under your Collector variable, select 'choose as G' to group your values by collector, and create your new color-coded graph.

**If you're happy with a particular visualization you created, click 'Save' at the top of your page. There is also an option to export your graph as a PDF or CSV, among other formats.*

**Tip: Once you've created a visualization, click on 'Traces' in the upper left corner to customize your graph. 'Traces' allows you to change bin size (making your view more or less specific), the sizes of markers on your graph, or to turn some groups on and off. For example, when we looked at the size of objects collected relative to collector, we clicked 'Traces' and 'Hide All,' and then added only the 'Mellon' group in order to focus more closely on that particular data set.*

**Tip: By clicking 'Themes,' you can change the colors and appearance of your graph.*

Second Exercise: With a sample of 1,500 records of British Art Sales from the Getty Provenance Index, which was too large a set for plot.ly, Matt illustrated the advantages of R, a free software for statistical computing and graphics, which is more capable of handling large data sets.

We worked with this data set to consider such questions as: the most popular and most lucrative times of the year for art sales, the variations of these patterns relative to the cost of artworks, and changes in these patterns over time.

When asked about the advantages of Plot.ly versus the Library of Congress' Viewshare Software, Matt explained that Viewshare has the capability to visualize data on a map and to add thumbnails of images to a data set and visualizations, which plot.ly does not. However, plot.ly allows for more detailed and customized visualizations than those created by Viewshare. Both are great and relatively user-friendly options, but the choice of which program to use ultimately depends on how you hope to visualize your data.