Project Overview

Purpose

In the final project, you will select, explore and describe different aspects of a data set using a combination of graphics, models and well-written text.

Instructions

In this project, you will select a data set and create graphics and models based on that data set. You will present your work in a presentation due at **11:59PM on Friday of Module 7**. You will summarize your work in a report **due at 11:59PM ET on Wednesday of Module 8.**

Data set selection

- **The data set needs to be of sufficient size** in terms of cases and variables to support a variety of analytic graphics and modeling methods addressed in class. Some students pick topics related to environment, health, crime or federal budgets. Some pick topics related to sports, popular music or movies. Some pick current interest topics.
- **Some data sets take a lot of pre-processing to be suitable for use** in analysis tools. Students vary in their skills sets that address pre-processing data. Students addressing the Boston marathon did a substantial amount of pre-processing to prepare the data. Some students struggle with their chosen data set and shift to another data set. For a few, too little time is left to produce a really good project. Keep the pre-processing in mind in the data selection process. Level of effort credit is given for substantial pre-processing work, but there must be a quality project presentation and paper.
- **If the intent is to produce maps as part of the project, then the data set needs to contain region names or identifiers and there must be access to region boundary files** and a way to use them. It is easier if the software already has the boundaries or the boundaries are available via working application example. The micromapST comes with US boundaries and makes a convenient choice for those interested in US states. Many students have other interests. The global administrate areas web site, GADM, has shape files for most nations. Just as the US has states, counties and census tracts, most nations have different levels of administrative regions. With R there are several ways to produce maps using shape files. There is an example in the R Graphic Cookbook. The micromap package (not micromapST) works with shapefiles, but there is a longer learning curve. The TCmaps java software also read shape files and also has many boundary files in working applications. The instructor can provide some guidance about boundary files.
- **The data set shapes the analysis.** If there is an interest in producing graphics for time series, then the data set needs to contain time series. The notion of data set here is quite general. It may contain pieces gathered from different places or put together as a set of files obtained from the same place.
- Once a data set is selected, **the challenge is often to limit the project scope** to stay within the page limit.

Graphics and Models

- **The project is to feature a variety of graphics and models based on the general methodology taught in class.** It many include just a few graphics or models using other methodology.
- **Providing variety is important.** There are many graphics from which to choose. For example, variants of dot plots include juxtaposed and superposed dot plots. The dots may have confidence

intervals. Values may be represented by bars or arrows. The family type and race dot plots were examples of conditioned dot plots. There can be multiway condition dot plots. There can be reference values. There are distribution plots, such as box, density and Q-Q plots. (The EDA plot includes all three.) There are correlation plots, scatterplots and scatterplot matrices. The scatterplots may show points or hexagon bins and may include smoothes with or without confidence bands. Graphics such as dendrogram can show the results of clustering. There can be glyph plots but often there are better alternatives. The graphics can be maps when data sets have a geospatial context. In this class linked, conditioned and comparative micromaps are often preferred to conventional choropleth maps since they often provide more context for interpretation and/or communication. Inclusion of conventional choropleth maps with good legends is an option.

- **Good projects typically include descriptive statistics and modeling results.** The class emphasized linear regression, random forest regression, and random forest classification. The model output and related graphics such as a regression diagnostics plots are candidates for inclusion.
- **Indicate the source of the data set and describe it briefly in text.**
- **Provide a coherent written description of the facets of the data addressed in the data exploration.**
- **The report should have at least one comment about every plot, map, or table included in the report.** A comment might be that there isn't much of pattern. There is no need to find a result of scientific or social importance. The exploration may address only a small part of large or complex data sets**.**
- **Avoid or limit the use of bulleted lists in the report.** The report is not a PowerPoint presentation.

Project page and time constraints

- The oral presentation is to be presented online within a **time limit of 7 minutes**.
- The final project report is limited to **10 pages**. The graphics may be a little smaller than ideal due to the limited page length.
- Do not spend time in the presentation or the space in the paper going over the four general graphics guidelines discussed in class. The class and the instructor should already know them.
- There may be an appendix that doesn't count against the page limit. The appendix may include R scripts and material related to the level of effort. The project report is due on the day of finals and serves as the class final.

Some general suggestions for writing the paper and/or the presentation outline

- **Start the paper by providing a quick answer to why this data set?** If a particular interest motivated the data set choice, it is good to express this. If you have questions you hoped to answer using the data, it is fine to mention this even if the data didn't provide an answer. Some people pick data sets because classmates will likely find them interesting.
- **Briefly describe the data set.** Think of a set of questions to answer. What does the data set address? Does it include geospatial and/or temporal data? What is the source of the data? How can data be obtained? Briefly, indicate the kind and amount pre-processing involved. (Extended material about pre-process can be put in an appendix.)
- **What are some of the variables that will be addressed?** You can start showing some univariate graphs and be off and running. That one organizational scheme is to start with simple graphics and lead up to more complex graphs and/or models. Diagnostics graphs typically follow models.
- **In telling a data exploration story, one way to think of it is as an expedition** (for example Lewis and Clarke expedition or a Darwin expedition) in which a researcher has gathered many samples and

returned. The researcher has studied the samples and selected a modest number to make a presentation to an audience. The researcher might decide to present the samples in the timeline of activities. There may be digressions to include something about obstacles encountered and addressed. This can convey more about the adventure or struggle. The conclusion might highlight a sample, share insights, or suggest a return expedition.

- The challenge is often to limit the examples based on the presentation time and page limits.
- See the bulleted lists below on variety and writing quality.
- See the appendix for overview of class methods.

Project Evaluation

The thoughtful investigation or description of a data set following the criteria below can warrant an A grade whether or not some exciting result is found. There are four broad evaluation areas indicated below: 1) level of effort, 2) adequate variety of methods taught in class, 3) graphics quality and 4) the writing quality.

In some cases, the majority of the project effort may be directed toward the process that precedes the exploration. For example, data cleaning can be a pain and take a long time. The report should explicitly communicate the level of effort when a substantial effort has been made to obtain and/or clean up the data for graphics and/or modeling. Do not assume the level of effort is obvious. Do not assume that effort by itself warrants a high grade. There must be a good project paper.

Level of effort areas

- Data gathering and preparation of data for analysis/graphics
- Development of a GUI to access data and facilitate graphics production
- Development and/or integration visual analytics
- Substantial adaptation of existing graphics or using graphics for new data
- Development of new graphics
- Creation of R functions or even an R package.

Adequate variety of methods taught in class

- Graphics should show variety, they may show one, two or more factors
    - One, two or more continuous variables
    - Time series
    - Spatial context and maps
    - Data distribution and functional relationship graphics
    - Model diagnostics
- Models
    - Can be as simple as fitting means
    - Multiple linear, logistic, and random forest regression
    - Random forest classification
    - Are not absolutely required for an A but are consider important

Graphics quality

- Provide the units of measure, etc.
- Follow the general graphics guidelines: making accurate comparisons, adding context for interpretation, and striving for simple appearance. The objective of engaging the reader can be

addressed using quality static graphics. However, a good way to engage the audience is using interactive and dynamics graphics, such as Shiny, CCmaps and TCmap in the presentations. The paper can show prints and include comments referencing the dynamic graphics shown in class.

- Avoid graphics deprecated in class. The deprecated methods in some cases go against common practice and perhaps even some perceptual studies.
    - Pie charts
    - Perspective bars without watermarks
    - Symbols plotted on the plot outline (exceptions will be made for ggplot facetted graphics because the axis limits cannot be controlled).
    - Axes with missing labels and/or units of measure

Writing Quality

- Papers are to be well written. Non-native English speakers are advised to get help with their writing. There are GMU resources for this. Native English speakers are also encouraged to help with their writing. Some students are already excellent writers, so need no help, but most people can benefit the input.
- Indicate the nature of the data set and its source
- Indicate the goals for the graphics and/or analysis (The goals can simply be exploration and description)
- Logical and/or systematic description and reasoning
- Indicate results or conclusions
- Clear labeling of figures and tables

Appendix: An overview of class methods

There are many ways to provide an adequate variety of examples illustrating methods taught in class. In terms of a continuous variable, the EDA plot includes density plot, a box plot and a normal Q-Q plot. Including an EDA plot provides a **start** to addressing the variety of objectives by have three different kinds of graphics. The density plot and the box plot are two kinds of univariate distribution graphics. The Q-Q plot is a bivariate plot comparing two distributions.

With many continuous variables a person could fill pages with the EDA plots but this would only show the same EDA multiple panel design over and over. As demonstrated in class, one can quickly scan many such plots. For the project, it is better to pick out one or two EDA plots to feature and describe. Comparison is the heart of graphics. We often want to compute and compare statistics for data indexed by factor levels (categorical variable classes). Examples visually compared height quantiles, densities, box plot statistics and means. We can make visual comparisons using juxtaposed or superposed graphics. The singer data used juxtaposition of voice part graphics because superposition with eight factor levels is problematic.

Project data sets have multivariate data and fairly often they involve a geospatial context and/or time. The class addressed bivariate patterns from the perspectives of data density and functional relationships using scatterplots and addressed multiple pairs of variables using scatterplot and correlation matrices. The trivariate analysis focused on using smoothes to call attention to functional relationships of form z=f(x, y). The crime data started with more the 100 variables and used combination of graphics and models to describe this data set.

The class and a class text presented a variety of micromaps. Micromaps and simple choropleth maps are within the class scope. TCmaps addresses maps and times series.

While times series were not stressed, times series graphics are within the scope of class. In the Visualizing Data Patterns with Micromaps text the example on page 73 shows state time series. The map could be omitted. Plots similar to this can be constructed by perceptual grouping times series based on a sorting variable such as the last value of each series, superposing series with each perceptual group and juxtaposing the groups. Both ggplot2 and lattice packages produce such graphics.

Clustering has many uses in data analytics. Cluster related graphs count toward variety. Multivariate glyphs were also taught but partly to indicate that they are often not very effective.

Data modeling represents a significant part of methods taught in class. Modeling began with the singer data and the fitting of the mean to height of each voice part. Many models produce fitted values that can be compared to the observed values. These can be readily shown in superposed plots. However, graphic comparisons are often best made by directly showing differences (or ratios). The class addressed the study of multiple linear residuals and regression diagnostics. The regression modeling methods include random forests and lasso regression. Random forests address classification and logistic regression model the probability of binary events. The class addressed variable selection using a variety of methods such as cross-validated all subsets selection in multiple linear regression, variable important in random forests and penalty function based shrinkage in Lasso regression.