BAE 590 R Coding for Data Management and Analysis – Final Project

Overview

The primary objectives of this project are to perform an exploratory data analysis (EDA) in R and communicate your findings via a written report. BAE 590 students will also summarize EDA findings in a 5- to 7-minute recorded presentation. Your EDA should attempt to answer at least one general driving question. Examples of general questions include:

- What are the most commonly applied fertilizer products in the US? How have commonly used products changed over time?
- Are there long-term changes in blue crab harvests in North Carolina? Does temperature explain inter-annual variability in blue crab harvests? Does rainfall?
- Where is pest damage in forests the worst across the US? How does pest damage vary by state?

Data: Your project dataset should include at least six variables and more than 500 rows. If you find a dataset that you *really* want to analyze and it does not have 500 rows, contact me and I will evaluate whether the dataset is adequate for the project. Note that you may have to join 2+ datasets in order to have a project dataset that has sufficient information to perform an EDA.

Outputs:

- Four to six quality data visualizations (no fewer than four, no more than six). Though you are limited to six visualizations total in your report, individual visualizations can include multiple panels (i.e. facets). One of your visualizations can be a geospatial visualization, such as a site map.
- At least one data summary table. There is no limit on the number of tabulated summaries you can include in your report.
- Results from a statistical analysis. The analysis you apply will vary depending on whether you are a BAE 495 or BAE 590 student:
    - BAE 495: Fit at least one simple or multiple linear regression model to evaluate relationships between one or multiple explanatory variables (i.e. independent variables) and one response variable (i.e. dependent variable). In the "Applying models" module, we will review how to fit linear regression models in R, so you will not be fitting regression models for the first time as part of this project.
    - BAE 590: Perform a statistical analysis or apply a model of your choosing. The statistical analysis *cannot* be simple linear regression. I expect that you will have to do some independent online searching to determine how to apply the analysis you have chosen. Examples of statistical analyses you may consider applying include hypothesis tests, principal component analysis, k-means clustering, etc.

Point breakdown

BAE 495

- Proposal: 5 points
- Report: 80 points
- Code: 15 points

BAE 590

- Proposal: 5 points
- Presentation: 10 points
- Report: 70 points
- Code: 15 points

Project proposal

The purpose of the project proposal is to initiate a conversation with me about your project idea so you can receive feedback prior to starting the project. Your proposal should be succinct. You are welcome to use bullet point formatting. The proposal must include the following:

- General question(s) driving your analysis.
- Data to be analyzed: the variables, data source, and any additional details you think I need to be able to understand the data you'll be analyzing. **Include the data file(s) as an attachment to your proposal submission.**
- BAE 590 students: The statistical analysis or model you plan to apply to your data.

Project report

- *Abstract – 10%*
  - Summarize your project report in one paragraph. Include key findings.
- *Background and driving questions – 15%*
  - Describe the background and rationale for your project in at least 10 sentences. This section should explain why the subject is interesting and important.
  - Provide at least one general question motivating your EDA. Note that the text you write to describe the question(s) does not count towards your 10-sentence minimum for the background description.
- *Dataset description – 10%*
  - Describe your data. Include the data source and/or methods used to collect the data; you do not need to provide detailed information on data collection, but should offer enough info so the reader can understand what information is captured in your dataset.
  - Summarize the variables in your dataset in a table. The table should include variable names, variable names as included in your data file (e.g., variable name = temperature; variable name in data file = T), and units.
- *Methods – 15%*
  - Outline key steps in your workflow; describe the steps with complete sentences, do not only use bullets/numbered lists. Include the following sub-sections:
    - — Data tidying and wrangling: what steps did you take to tidy and organize your data?
    - — Visualizations: which visualizations did you create, and why?
    - — Statistical summaries: how did you summarize your data, and why?
    - — Statistical analysis: what was the statistical method you applied, and why?
- *Results and Discussion – 50%*
  - Include the four data visualizations and data summary you produced from your EDA. The figures should adhere to the key principles of data visualization.
  - Include figure and table captions. Note: table captions are provided above tables, whereas figure captions are provided below figures.
  - Describe that insight you gained from your EDA. In particular, **be sure to answer your driving questions (general and specific)**. Or, if you feel you do not have enough information in your data to answer your driving questions, explain why and describe what additional data would

be needed to answer your driving questions. This section should be a minimum of 4 paragraphs.

Presentation (*BAE 590 students only*)

Record a 5-7 presentation summarizing your project. You may not have enough time to present all of your figures and data summaries, so only include key findings. You can record your presentation as an audio + screen recording (e.g. with QuickTime or similar free program). Alternatively, using your phone or other recording device, you can record a video of yourself delivering the presentation.

R script(s)

Your entire analysis should be entirely reproducible, meaning that all steps of your analysis must be performed in R. The script(s) you prepare to perform your EDA and generate outputs should be organized/annotated and have code headers. The codes must be submitted for the project to be graded. The scripts will be graded based on organization, not specific content.