GDAA 1001 - Fundamentals of Spatial Data Analytics

Assignment #2 - Exploratory Data Analysis

Due: 14/11/2021 at 11:59pm | Value: 20% of Final Grade

Introduction

For this assignment, you are tasked with carrying out Exploratory Data Analysis (EDA) on a dataset of your choice. You will carry out a range of tasks with the aim of gaining further understanding of your data. You will:

- Describe your data using summary statistics
- Examine variation within and covariation between your variables
- Visualize distributions of variables using density plots, histograms, bar graphs and box plots
- Visualize the relationships between variables using scatterplots and heatmaps
- Discuss your results

Below is a list of tasks to be completed. These tasks comprise the basic elements of EDA, and serve to help identify the key features of a dataset and important relationships between specific variables. A rubric is provided at the end of the document.

Resources

Exploratory Data Analysis can include a wide range of processes, depending on the complexity of the dataset involved. A good overview of the EDA process is found in Chapter 7 of *R for Data Science* by Wickham & Grolemund (https://r4ds.had.co.nz/exploratory-data-analysis.html).

Data Selection

You are responsible for selecting a suitable dataset for this assignment. The parameters for selecting an appropriate dataset are as follows:

- It should be an external dataset (i.e., not a built-in dataset within R or its extensions)
- It can be either spatial or non-spatial
- It should contain at least 100 observations/rows and 5 variables/columns
- At least 2 variables should be factors/categorical (note: you can convert a numeric variable into a categorical data if necessary)

Locating datasets can be challenging. Most cities/municipalities offer Open Data portals that link to many freely available datasets, including some with spatial attributes (e.g., Halifax Open Data: https://catalogue-hrm.opendata.arcgis.com/). Another good resource for sample datasets is Kaggle (https://www.kaggle.com/datasets).

Tasks

Produce a report with the following sections:

- Introduction
- Data Selection and Preparation
- Data Summary
- Exploration of Variation
- Exploration of Covariation
- Discussion

Introduction

If generating a PDF, your report should have a cover page including a title, the submission date, your name and course information. You should also include a table of contents and list of figures. If generating an HTML report or using a Jupyter Notebook, ensure that the above are located at the beginning of the document.

In your introductory section, describe the purpose of the assignment, introduce your dataset, and outline some general questions about your data. What do you expect to find in your EDA with regards to the distributions and correlations within and between your variables? Pose some general questions prior to your analysis, and then revisit these questions later in your discussion, after you have explored your data thoroughly.

Data Selection and Preparation

Provide a brief description of your data selection process. Where did you find your dataset? Why did you choose this dataset?

Describe your procedures for importing, cleaning and tidying your data. For example, did you have any missing values? If so, how did you handle these missing data? Did you have to convert any variables to factors? Any details relating to data preparation should be included here.

Data Summary

Include a brief summary of your data using the outputs from `str` and `summary`. Format your outputs so that they appears a tables, and not just console outputs. A good way to output data stored as either a `data.frame` or a `tibble` is to use the `knitr::kable()` function.

Exploration of Variation

Provide a thorough examination of variation within your variables. Use `ggplot2` to generate your graphical outputs. This section should include:

- Bar charts for your factors/categorical variables
- Histograms, density plots, and boxplots for your numeric variables

Each plot should be formatted appropriately, with accurate axis labels. Conclude your section with a very brief description of variation based on your observations of these plots.

Note: Your dataset may have many numeric and categorical variables. You do not need to include an exhaustive set of plots of each variable. Rather, make sure you include at least one example of each type of plot (histogram, density plot, boxplot, bar chart), or perhaps a couple of each. You can decide which to include.

Exploration of Covariation

Provide a thorough examination of covariation between your variables. Use `ggplot2` to generate your graphical outputs. This section should include:

- At least one scatterplot and one scatterplot matrix
- At least one heatmap

Each plot should be formatted appropriately and have accurate axis labels and legend labels. Include a brief description of covariation based on your observations of these plots.

Note: Similarly, you do not need to include individual scatterplots or heatmaps of all possible combinations of variables. Rather, pick an interesting example or two for your numeric and categorical variables and plot those using scatterplots and heatmaps. If you have many numeric variables, you should also include a scatterplot matrix (e.g., four or more numeric variables).

Discussion

Conclude your report with a brief discussion of the results of your EDA. Revisit the questions initially outlined in your introductory section. Has your analysis revealed any interesting or unexpected patterns in your data, whether in terms of distributions within given variables or correlations between sets of variables?

Describe how visualizing your data has helped you to gain a better understanding of your selected dataset, and identify which combinations of variables you think would be most important in a future modeling exercise carried out on your data.

Submission Instructions

For this assignment, you are required to submit:

- Either a PDF or HTML report generated from an .Rmd file, or a Jupyter Notebook
- If opting for a PDF or HTML report, the .Rmd file used to generate your report should also be submitted
- A copy of your dataset (uploaded to Brightspace if possible; otherwise, provide a link to the raw data)

Submissions should be placed in the *Assignment #2 – Exploratory Data Analysis* folder on Brightspace by the specified deadline. Late submissions will be subject to a 10% per day penalty.

Evaluation (Rubric)

Assignment #2 is valued at 20% of your final grade. The following rubric scores your assignment out of 30 possible points. Use this rubric alongside the task list above as a guide to successfully completing this assignemnt.

	5 Points	**2.5 Points**	**0 Points**
Formatting & Organization	Excellent formatting and organization; clearly defined sections with appropriate headings; no spelling/grammatical errors	Good formatting and organization; some appropriate section headings; a few spelling/grammatical errors	Poor formatting and organization; few or no sections/headings; many spelling/grammatical errors
RMarkdown File	.Rmd file works on my machine; YAML is appropriate; code chunks are formatted correctly; thorough and detailed comments in your code	.Rmd file works on my machine; YAML is appropriate; code chunks are generally formatted correctly; some comments are provided in your code	.Rmd file works with some issues, or does not work properly at all; code chunks poorly formatted; limited or no comments in your code
Data Selection & Preparation	Choice of data is appropriate with many observations and variables; correct mix of variable types; data preparation is thoroughly and clear	Choice of data is mainly appropriate with many observations and variables; better mix of variable types required; data preparation is somewhat thorough, but requires further detail	Choice of data not appropriate; too few observations or variables; improper mix of variable types; no focus on data preparation procedures
Exploration of Variation	Detailed description of variation within dataset; all variables are plotted using appropriate plot types	Somewhat detailed description of variation; not all variables are plotted, or misuse of plot types	Superficial description of variation; improper or missing plots
Exploration of Covariation	Detailed description of covariation within dataset; all variables are plotted using appropriate plot types	Somewhat detailed description of covariation; not all variables are plotted, or misuse of plot types	Superficial description of covariation; improper or missing plots
Discussion and Write-Ups	Thorough and insightful analysis with detailed discussion of patterns and trends; revisit introductory questions	Basic analysis with some discussion of patterns and trends; introductory questions only cursorily revisited	Lacking write-ups with little to no analysis