Lecture 1, MTH 400, Fall 2024

Intro to EDA
Intro to the Tidyverse

Exploratory Data Analysis (EDA) is the process of analyzing and summarizing data sets in order to gain insights and understand the underlying patterns, relationships, and trends within the data. EDA typically involves using statistical methods and data visualization tools to explore the data, identify any outliers or anomalies, and uncover any underlying structure or relationships within the data.

The goal of EDA is to gain a deep understanding of the data, without making any assumptions or preconceived notions about what the data may reveal. This can involve looking at the distribution of data, identifying any correlations between variables, and understanding the central tendencies of the data. EDA can be conducted on both small and large data sets, and can be used in a wide range of applications, from scientific research to business intelligence.

Some of the techniques used in EDA include histograms, scatter plots, box plots, correlation matrices, and heat maps. These techniques can help to identify any patterns or trends within the data, and can also help to identify any outliers or anomalies that may need to be investigated further. Overall, EDA is an important step in the data analysis process, as it can help to identify any potential issues or areas of interest in the data, which can then be further explored and analyzed using more sophisticated statistical techniques.

Graphical methods can be used to explore data in a variety of ways. Here are some of the common ways:

1. Visualizing the distribution of data: Graphical methods can be used to plot the distribution of a variable, such as a histogram or a density plot. These visualizations can help to identify the range of values, any skewness or symmetry, and the overall shape of the distribution.
2. Identifying relationships between variables: Graphical methods can be used to plot the relationship between two variables, such as a scatter plot or a line plot. These visualizations can help to identify any patterns or trends in the data, as well as any outliers or influential observations.
3. Comparing groups or categories: Graphical methods can be used to compare the distribution of a variable across different groups or categories, such as a bar plot or a box plot. These visualizations can help to identify any differences or similarities in the data across groups.
4. Visualizing time series data: Graphical methods can be used to plot time series data, such as a line plot or a stacked area plot. These visualizations can help to identify any trends, seasonal patterns, or changes over time.

Overall, graphical methods can be a powerful tool for exploring data, as they can help to identify patterns and relationships that may not be apparent from looking at the data alone. They can also be used to communicate insights and findings to others in a clear and intuitive way.

Tabular methods for summarizing data for both numerical and categorical data are also important and can provide key insights.

The tidyverse is a collection of packages in R designed for data science and analysis. Here are some of the tools available in the tidyverse for exploring data:

1. ggplot2: ggplot2 is a package for creating data visualizations, including scatter plots, line plots, bar plots, and more. It allows for customization of colors, labels, and other aesthetic features, and can be used to create complex visualizations with multiple layers.

2. dplyr: dplyr is a package for data manipulation, including filtering, sorting, grouping, and summarizing data. It allows for easy and intuitive manipulation of data frames, and can be used to quickly explore and summarize data sets.
3. tidyr: tidyr is a package for data reshaping, including converting data between wide and long formats, and filling missing values. It can be used to clean and prepare data sets for analysis and visualization.
4. readr: readr is a package for reading in data from various file formats, including csv, tsv, and Excel files. It can be used to quickly import data sets for analysis.
5. purrr: purrr is a package for functional programming, including mapping, filtering, and reducing data. It can be used to apply functions to data sets and perform complex operations on them.

Overall, these tools can be used in combination to explore and analyze data sets, and can be particularly useful for large and complex data sets that require efficient and flexible methods of manipulation and visualization.

In addition to the packages within the tidyverse, there are several other packages in R that can be used for exploratory data analysis. Here are some popular ones and their main uses:

- base R: The base R package comes with R by default and provides a wide range of functions for data manipulation, summary statistics, and basic plotting. It includes functions like summary(), plot(), hist(), and cor() that can be used for basic exploratory data analysis.
- data.table: data.table is a package that offers fast and efficient data manipulation operations on large data sets. It provides an alternative syntax to base R and can handle millions or billions of rows of data with ease. It is particularly useful for tasks such as filtering, grouping, and aggregating data.
- lattice: lattice is a package for creating trellis plots, which are a type of multi-panel data visualization. It allows for the creation of conditioned plots, such as scatter plots, bar plots, and box plots, with separate panels for different subsets or categories of data. It is especially useful for exploring relationships between variables in multivariate data.
- ggvis: ggvis is a package that extends the grammar of graphics framework from ggplot2 to interactive visualizations. It allows for the creation of interactive plots that can be explored dynamically, such as adding tooltips, zooming, and panning. It is suitable for creating interactive visualizations for exploratory purposes.
- plotly: plotly is an interactive plotting package that can generate web-based, interactive visualizations. It allows for the creation of interactive plots with features like zooming, panning, tooltips, and hover effects. It can be particularly useful for creating exploratory visualizations that can be shared and explored online.
- corrplot: corrplot is a package for visualizing correlation matrices. It provides various methods to display correlation coefficients between variables, such as color-coded matrices, scatter plots, and network plots. It is useful for exploring relationships and dependencies between multiple variables.

These are just a few examples of packages available in R for exploratory data analysis. Depending on your specific needs and the nature of your data, there are many other packages that can be used to enhance your data exploration process in R.

Exploratory Data Analysis (EDA) involves various aspects beyond data visualization. Here are some additional packages in R that can be useful for other aspects of EDA:

- Hmisc: Hmisc is a package that provides functions for data manipulation, descriptive statistics, and advanced statistical analysis. It includes functions for handling missing values, creating frequency tables, calculating summary statistics, and conducting hypothesis tests.
- psych: The psych package offers a range of functions for psychological and psychometric analysis. It includes tools for factor analysis, clustering, reliability analysis, and descriptive statistics. These functions can be helpful for understanding the underlying structure of data and exploring relationships between variables.
- caret: caret (Classification And REgression Training) is a package that focuses on machine learning and predictive modeling. It provides functions for data preprocessing, feature selection, model training, and evaluation. While primarily used for modeling, it can also aid in EDA by identifying important features and assessing the performance of different models.
- FactoMineR: FactoMineR is a package for multivariate exploratory data analysis and dimensionality reduction techniques. It includes functions for principal component analysis (PCA), correspondence analysis, and clustering. These methods can help to uncover patterns, similarities, and differences in high-dimensional data.
- missForest: The missForest package is specifically designed for imputing missing values in data sets. It provides functions for imputing missing values using a random forest algorithm. This can be valuable in EDA when dealing with incomplete data, allowing for a more comprehensive analysis.
- Amalia: Amalia is a package for anomaly detection and outlier analysis. It offers various algorithms and techniques to identify unusual observations or patterns in data sets. These methods can be used to explore and investigate potential outliers or anomalies in the data during EDA.

These packages provide additional functionality beyond data visualization and can assist with tasks such as data manipulation, summary statistics, statistical analysis, dimensionality reduction, missing value imputation, and anomaly detection, all of which are integral to a comprehensive EDA process.

Clustering methods can also be used for data exploration. While we will not be learning clustering methods directly in this course (they are covered in CSC 400 Data Mining), if you are familiar with these techniques or wish to learn them on your own, you are also free to use these methods.

Resources:

1. https://r4ds.had.co.nz/exploratory-data-analysis.html
2. https://towardsdatascience.com/tidyverse-vs-base-r-how-to-choose-the-best-framework-for-you-29b702bdb384
3. https://www.listendata.com/2014/06/data-exploration-using-r.html
4. https://bookdown.org/palmjulia/r_intro_script/data-exploration.html
5. https://www.geeksforgeeks.org/exploratory-data-analysis-in-r-programming/
6. https://geanders.github.io/RProgrammingForResearch/exploring-data-1.html
7. https://www.analyticsvidhya.com/blog/2015/04/comprehensive-guide-data-exploration-r/
8. https://forum.posit.co/t/data-exploration-in-r/128336
9. https://r4ds.had.co.nz/explore-intro.html