STAT 471: Introduction to R Programming Lecture

Lecture 7: Data Visualization Continued and Exploratory Data Analysis (EDA)

Cory Suzuki

Department of Mathematics & Statistics
California State University, Long Beach

1 October 2025

Overview

1. Data Visualizations and EDA

2. Basic EDA

3. Advanced EDA

Data Visualizations

The past 6 weeks have gone over foundational material such as the basics of R datatypes, programming structures, some helpful algorithms, and the basic types of data visualizations. We've taken a look at:

- Scatterplots
- Histograms
- Bar Charts
- Boxplots
- Scatterplot Matrices

Exploratory Data Analysis

Now, we can apply these different types of visualizations and built-in R functions to perform Exploratory Data Analysis (EDA). Think of this as a preliminary analysis prior to building statistical models. The EDA pipeline consists of:

- Wrangling/Collecting the Data
- Cleaning the Data and Creating a Dataset (includes any necessary data transformations)
- Summarizing your data with Summary Statistics (Mean, median, min, max, quartiles, etc.)
- Investigating the distributions of your feature columns
- Finding any correlations or relationships between your features of your dataset

What Else Can Be Done with EDA?

- Cluster Analysis: For categorical features, we may want to find how observations are classified and visualize observations as groups or clusters
- Dimensionality Reduction: In multivariate statistical analysis cases, we may want to reduce the number of features we are analyzing while preserving the original information of the dataset (high dimensional data to low dimensional data)
- Cluster Analysis and Dimensionality Reduction often use Unsupervised Machine Learning Models to draw conclusions about how we can further our analyses
- Unsupervised ML models work with unlabeled responses in our data and doesn't depend on predefined categories in categorical variables. Supervised models do work with labeled data (linear regression is an example of a supervised model)

Summarizing Data

Summarizing data usually entails finding summary statistics for the features within our dataset. We can use summary() within R to retrieve that information for us. Primarily, we want to get insights about our data such as:

- mean
- quartiles
- median
- minimum
- maximum

This allows us to describe our data as a preliminary check. It is also a good idea to display the datatypes of your features and counts for categorical features.

Distributional Checks

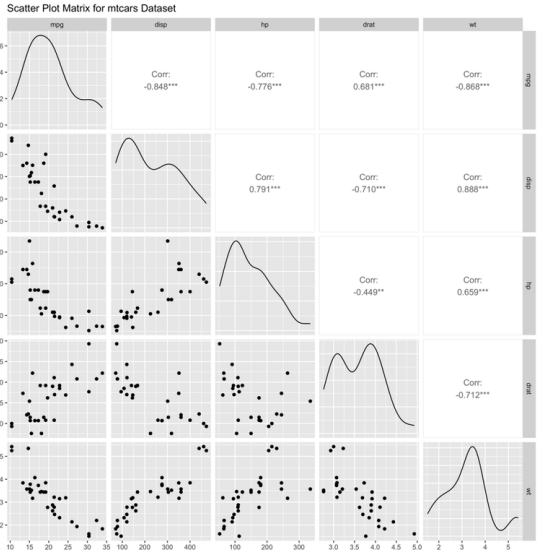
Histograms and bar charts (for numerical and categorical data respectively) help in checking the distributions of your features. Doing this also allows you to see whether you can apply the appropriate imputation techniques for missing data (recall median imputation for skewed data and mean imputation for normal data)

Scatterplot Matrices and Heatmaps are your Best Friends for Correlations

As we saw in the lecture on foundational data visualization techniques, scatterplot matrices allow you to gather insights about the correlations between your features. Heatmaps are also a good way to do this (I owe you an example today) since they show you correlations by their strength.

Brighter colors show weaker correlations, darker colors show stronger correlations between your features.

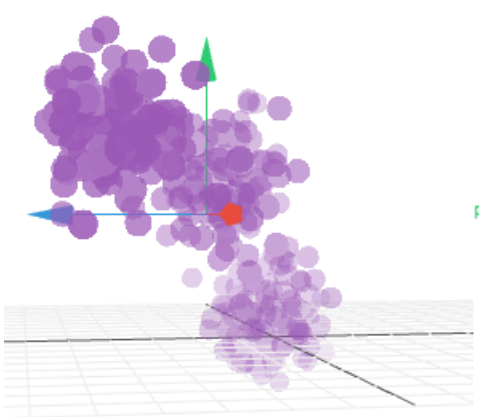
Scatterplot Matrix Example

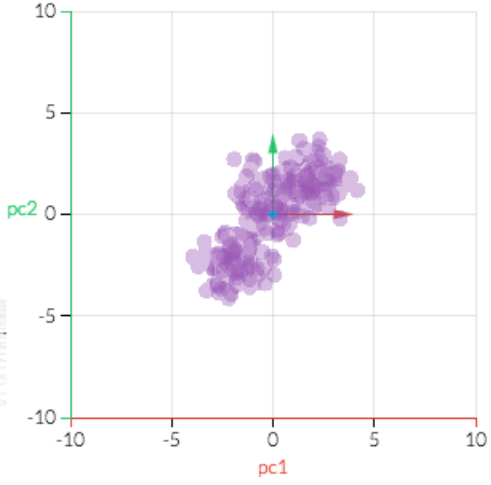


Dimensionality Reduction via Principal Components Analysis (PCA)

- Goal: Reduce the number of dimensions/features in a dataset while preserving the overall information of the features
- This method centers the features by subtracting the feature's mean from each value
- A covariance matrix that contains all the computed covariances and variances between each of the features is then computed
- Using linear algebra, the eigenvalues and eigenvectors of that covariance matrix is computed, where each eigenvector will represent the directions of maximum variance for each feature
- The first principal component is a linear combination of the features that explain the most variance, while the second component is a linear combination of the features that explain the least variance

Example of PCA





Disadvantages of PCA

- While we can explain the linear combinations of features that give us the maximum and minimum variance in our data, the principal components themselves are hard to interpret

- This is a linear dimensionality reduction method since it requires eigenvalues to be calculated, so the method may fail with nonlinear datasets

- For nonlinear relationships, you may need to use nonlinear dimensionality reduction via t-Stochastic Neighborhood Embeddings or t-SNE(STAT 576: Data Informatics/Unsupervised ML)

- In t-SNE, high-dimensional data is modeled with Gaussian distributions and low-dimensional data is modeled with t-distributions, the minimized divergence (difference between probability distributions) between these distributions results in a new, reduced dataset

Next Time...

- SQL and Database Manipulation in R
- Midterm Part A Study Guide coming soon! (to be released next Wednesday)

Announcements

Programming Assignment 2 is due this Saturday at 11:59pm!