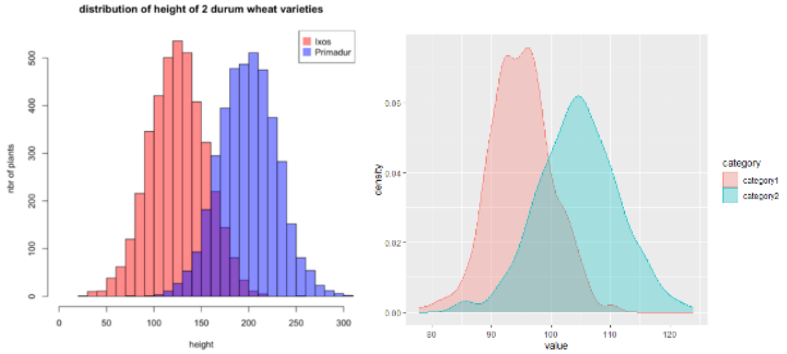
Lecture 3, MTH 400, Fall 2024

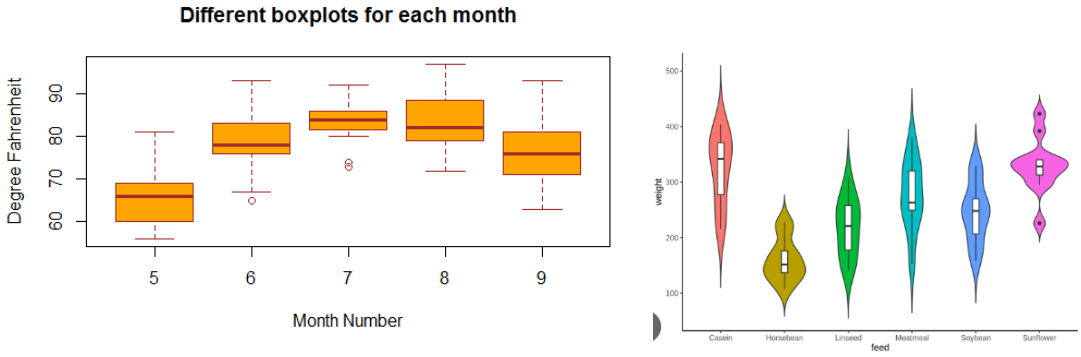
Visualizing and Summarizing Numerical Data

When conducting exploratory data analysis (EDA) on numerical data, the goal is to understand the distribution, central tendency, variability, relationships, and other important characteristics of the data. Here are some common exploratory analysis techniques used for numerical data:

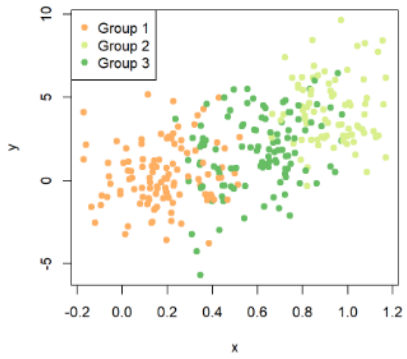
1. Descriptive Statistics: Calculate basic summary statistics such as mean, median, mode, standard deviation, variance, minimum, maximum, and quartiles. These statistics provide an overview of the central tendency, spread, and shape of the data distribution.
2. Histograms: Create histograms to visualize the frequency distribution of the numerical data. Histograms help identify the shape of the distribution, presence of outliers, and skewness. Variations can include a variety of comparative options and density plots.



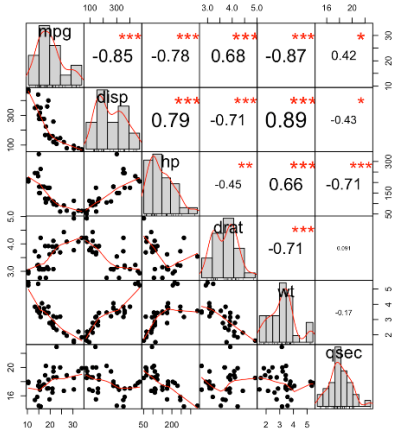
3. Boxplots: Construct boxplots (also known as box-and-whisker plots) to display the distributional properties of the data, including median, quartiles, and outliers. Boxplots provide a visual summary of the central tendency, spread, and skewness of the data. Variations can include swarm plots, strip plots and violin plots.



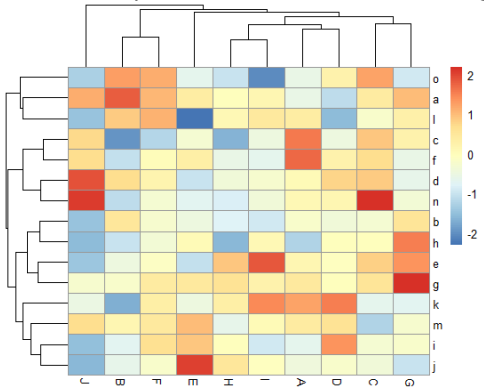
4. Scatter Plots: Generate scatter plots to visualize the relationship between two numerical variables. Scatter plots help identify patterns, trends, and potential correlations or associations between the variables. Clustering and regression information can be added to these plots.



5. Correlation Analysis: Calculate correlation coefficients (such as Pearson's correlation coefficient) to measure the strength and direction of the linear relationship between two numerical variables. Correlation analysis helps identify variables that are positively or negatively correlated.

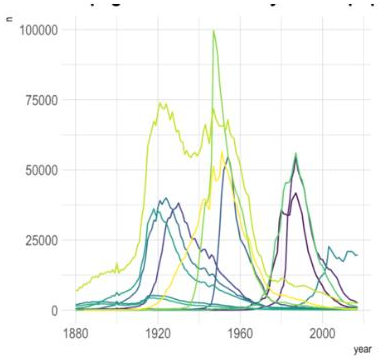


6. Heatmaps: Use heatmaps to visualize the correlation matrix of multiple numerical variables. Heatmaps provide a graphical representation of the pairwise correlations and help identify clusters or patterns of association among variables.



7. Outlier Detection: Identify outliers in the numerical data using techniques such as z-scores, modified z-scores, or Tukey's fences. Outliers are extreme values that deviate significantly from the rest of the data and may warrant further investigation.
8. Distribution Fitting: Fit probability distributions (e.g., Gaussian, exponential, etc.) to the data and assess how well the data aligns with different distributional assumptions. This helps understand the underlying distribution and assess goodness of fit.

9. Time Series Analysis: If the numerical data is time-dependent, apply time series analysis techniques such as decomposition, autocorrelation, and seasonality analysis. Time series analysis helps uncover patterns, trends, and cyclic behavior within the data.



10. Hypothesis Testing: Conduct hypothesis tests, such as t-tests or ANOVA, to compare numerical variables across different groups or conditions. Hypothesis testing helps determine if there are statistically significant differences in the means or variances of the groups.

These are some common techniques used in exploratory data analysis for numerical data. The choice of techniques depends on the specific characteristics of the data and the research questions or objectives. It's important to apply appropriate statistical measures and visualizations to gain insights into the distributional properties, relationships, and patterns within the numerical data.

Resources:

1. https://book.stat420.org/summarizing-data.html
2. https://openintro-ims.netlify.app/explore-numerical
3. https://campus.datacamp.com/courses/exploratory-data-analysis-in-r/exploring-numerical-data?ex=1
4. https://www.stat.uchicago.edu/~yibi/s220/labs/lab02.html
5. https://rpubs.com/Arundharshini/733988
6. https://r-graph-gallery.com/histogram.html
7. https://www.datamentor.io/r-programming/box-plot
8. https://r-charts.com/correlation/scatter-plot-group/
9. http://www.sthda.com/english/wiki/correlation-analyses-in-r
10. https://www.r-bloggers.com/2022/05/how-to-draw-heatmap-in-r-quick-and-easy-way/
11. https://r-graph-gallery.com/line-plot.html
12. https://www.geeksforgeeks.org/multiple-density-plots-and-coloring-by-variable-with-ggplot2-in-r/
13. https://jtr13.github.io/cc20/introduction-to-violin-plots.html