**Chapter 1: Exploratory Data Analysis (EDA):**

**Learning Objectives:**

1. What is EDA and Why is it important?
2. Numeric and Visual tools to perform EDA
3. How to conduct EDA using R and Python – What are the benefits and disadvantages?

**Outline:**

**What is exploratory Data Analysis (EDA)? Why is it important?**

**What Numeric and Visualization tools are used to perform EDA? – Brief discussion on topics 1 – 8 below**

**Conducting EDA in R and Python – What is essential?**

* How to Read data from CSV files into R and Python
* Discussion on libraries necessary in R and Python for performing EDA

**How to Perform EDA:**

1. Descriptive Summary Statistics – Quick Peek at data ranges and Data validation
2. Preliminary assessment of data characteristics using Summary Statistics
3. Correlation with Output
   1. Pearsons, Spearman Rank Correlation, Tetrachoric and Biserial correlations
   2. <https://personality-project.org/r/html/tetrachor.html>
4. Correlation among inputs
   1. Why is this a problem?
5. Visualization of Relationships (Why correlation analysis is better done before this?)
   1. X-Y plots
   2. Matrix of X-Y Plots
6. Outlier detection and Summary Measures
   1. Box plots
   2. Violin Plots
7. Distributional Evaluation
   1. Histograms
      1. Binning Rules (Scott, Freedman-Diaconis, etc)
   2. Plotting Position formulas
      1. Weibull, Gringorten, Hazen, etc.
   3. Kernel Density Estimation (KDE)
8. Bivariate Relationships
   1. Heat maps (summary across categorical variables)
   2. Contour Plots

**Closing Remarks**

What are the advantages of R in EDA?

What are the advantages of Python in EDA?