DATA EXPLORATION

Outline

- Inspect
- Subset
- Edit
- Summarize
- Visualize

Data Exploration

- Data exploration is a crucial first step in any machine learning workflow.
- Helps data scientists understand the
  - *structure,*
  - *quality, and*
  - *patterns in the data.*

Inspect Data

- Examine data format
  - *Plain text (e.g., csv, .txt), spreadsheet (e.g.,.xlsx), columnar/binary (e.g., parquet, feather, pickle), or statistical format (e.g., SPSS, SAS)*
- Examine data to gain insight on
  - *size*
  - *structure*
  - *data types*
  - *completeness*
  - *consistency*
  - *encoding*

Subset

Machine learning problems often involve

- selecting relevant features
- using observations that meet analysis requirements
  - *For e.g.: individuals who have visited the website in the last one month, a random sample of 70% of data*

Edit Data

- This is essential for making datasets cleaner, more interpretable, and analysis-ready.
- Change variable names to align with analysis needs.
  - *For instance, names that are too long, uninformative or do not comply with naming rules*
- Derive new variables from existing variables
- Identify missing values and replace them with an identifier or a value
- Change data types
  - *If data was parsed incorrectly; accommodate greater precision; reduce memory requirement*
- Sort
- Delete rows or columns

Summarize

- Eyeballing even a moderate sized dataset is unlikely to yield useful insights
- Numerical summaries can yield insights independent of the actual size of the data
- For numeric variables
  - *Measures of central tendency: Mean, Median, Mode*
  - *Measures of dispersion: Standard Deviation, Variance, Range*
  - *Shape of the distribution: Skewness, Kurtosis*
- For categorical variables
  - *Frequency or Proportion*
  - *Cross-tabulations*

Visualize

- Early in the analysis pipeline, visualizing data can aid in
  - *forming an understanding of the data,*
  - *identifying trends, and*
  - *spotting anomalies*
- Common chart types
  - *Histogram: Distribution of variable*
  - *Scatterplot: Relationship between a pair of variables*
  - *Line plot: Trend over time*
  - *Bar chart: Compare a numeric variable across levels of a categorical variable*

Conclusion

- In this module, we examined the following data exploration steps
  - *Inspect*
  - *Subset*
  - *Edit*
  - *Summarize, and*
  - *Visualize Data*