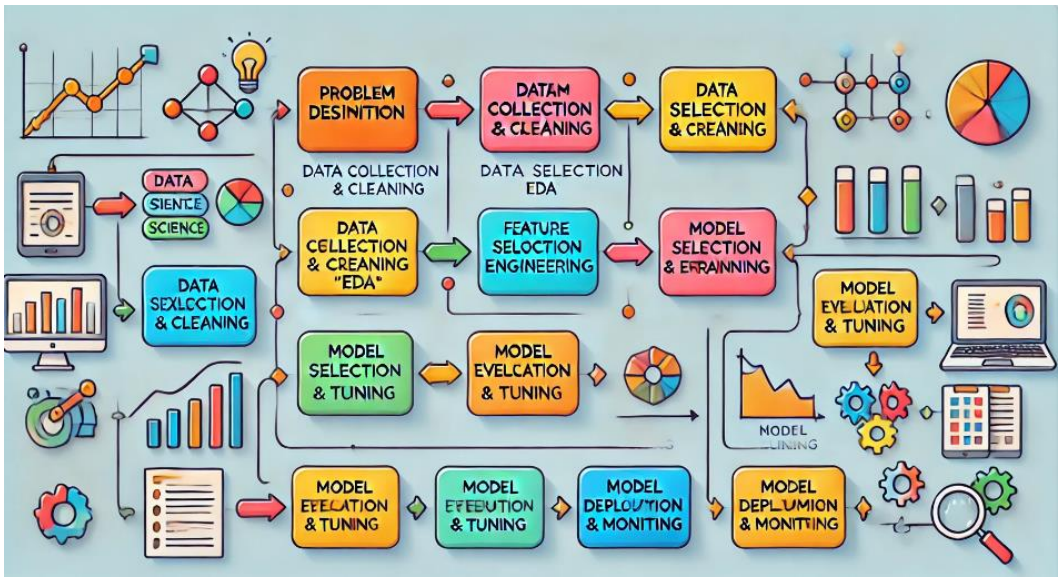
Computer Science Department
CS660 – Mathematical Foundations of Analytics (CRN# 22921)
Spring 2025

Project #1 / Due 27-Feb-2025

Exploratory Data Analysis (EDA) is a crucial step in any data science or machine learning project. It serves multiple purposes:

1. **Understanding the Data**: EDA helps in getting a better understanding of the data, including its distribution, relationships, and patterns. It allows you to get a sense of the data's underlying structure, which is essential for making informed decisions in later stages.
2. **Data Cleaning**: Through EDA, you can identify missing values, outliers, and inconsistencies in the data. Addressing these issues early on ensures the quality and reliability of the data.
3. **Hypothesis Generation**: EDA helps in generating hypotheses about the data. By visually exploring the data, you can identify potential relationships and correlations that may not be immediately obvious.
4. **Feature Selection and Engineering**: EDA aids in selecting the most relevant features and engineering new ones that could improve model performance. It helps in understanding which variables have the most predictive power.
5. **Model Selection**: Understanding the data through EDA allows you to choose the most appropriate algorithms for your problem. For instance, if your data is not linearly separable, you might consider non-linear models.
6. **Risk Mitigation**: EDA helps in identifying potential pitfalls and risks, such as data leakage or biased data, that could impact the model's performance.



For this data analysis project, we'll be using the following dataset:
Data Source: Electric Vehicle Population Data in Washington State, USA, available at https://data.gov/
Data Name: EV_Population_WA_Data.csv (already uploaded into the portal)
Data Description: This CSV file contains 200,048 rows, each representing an electric vehicle. The dataset includes the following columns: VIN, County, City, State, Postal Code, Model Year, Make, Model, Electric Vehicle Type, Clean Alternative Fuel Vehicle (CAFV) Eligibility, Electric Range, Base MSRP, Legislative District, DOL Vehicle ID, Vehicle Location, Electric Utility, and 2020 Census Tract.

Project Purpose:
The goal of this project is to create a Notebook (Jupyter, Colab) and become familiar with the Pandas and Matplotlib/Seaborn/Plotly libraries.

Questions to be Answered:
1. Which car manufacturers are the most commonly used for EVs in Washington?
2. What are the highest and lowest electric ranges in this dataset, and which car makers and models do they correspond to?
3. Is the maximum electric range value unique? If not, which cars share this range?
4. Is the minimum electric range value unique? If not, which cars share this range?
5. How does the electric range vary between car makers and between models?
6. Which are the top 5 cities adopting EVs?
7. How does the EV adoption rate vary among car makers over the years?
8. Is there a correlation between the electric range and the city of an EV?
9. Which county has the greatest variety of EV car models?

EDA Approach / Recommended, Best Practices
Here is a detailed list of steps for **conducting Exploratory Data Analysis (EDA)** on a given dataset:

1. Understand the Dataset Context
- Objective Clarification: Define the purpose of the analysis and the questions you aim to answer.
- Data Source Identification: Determine the source of the data, its collection methods, and any relevant background information.

2. Import Libraries and Load Data
- Import Necessary Libraries: Load libraries such as pandas, numpy, matplotlib, seaborn, etc.
- Load the Dataset: Load the dataset into a DataFrame using pandas (assume using Python).

3. Initial Data Inspection
- View Data Structure: Use functions like .head(), .tail(), .info(), and .describe() to get a sense of the data structure, types, and summary statistics.
- Check Dimensions: Identify the number of rows and columns using .shape().
- Identify Missing Values: Use .isnull().sum() to check for missing values.

4. **Data Cleaning**
   - Handle Missing Data:
     - Impute Missing Values: Fill missing values using strategies like mean, median, or mode imputation, or domain-specific methods.
     - Remove Missing Values: Drop rows or columns with missing data if appropriate.
   - Handle Outliers:
     - Detect Outliers: Use visualizations (e.g., box plots) or statistical methods (e.g., Z-scores) to identify outliers.
     - Treat Outliers: Depending on the context, either remove, transform, or cap outliers.
   - Correct Data Types: Ensure that each feature has the correct data type (e.g., convert columns to categorical, datetime, etc.).
   - Handle Duplicates: Check for and remove duplicate records.

5. Univariate Analysis
   - Summary Statistics: Review measures of central tendency (mean, median) and dispersion (standard deviation, variance).
   - Visualize Distributions: Use histograms, box plots, and bar charts to understand the distribution of individual variables.

6. Bivariate Analysis
   - Correlation Analysis: Calculate correlation coefficients (e.g., Pearson, Spearman) to understand relationships between numerical variables.
   - Cross-tabulation: Analyze relationships between categorical variables using crosstabs.
   - Visualize Relationships:
     - Scatter Plots: For numerical variables.
     - Box Plots and Violin Plots: To compare distributions across categories.
     - Heatmaps: To visualize correlations.

7. Multivariate Analysis
   - Pairplot/Scatterplot Matrix: Use to visualize relationships between multiple numerical variables.
   - Multivariate Statistics: Explore techniques like Principal Component Analysis (PCA) for dimensionality reduction.
   - Advanced Visualizations: Consider using more complex visualizations like pairwise correlation heatmaps or 3D plots.

8. Feature Engineering
   - Create New Features: Based on domain knowledge or interaction terms.
   - Feature Transformation: Normalize or standardize features, apply log transformations, etc.
   - Encoding Categorical Variables: Convert categorical variables to numerical using techniques like one-hot encoding, label encoding, or frequency encoding.

9. Handle Imbalanced Data (If Applicable)
   - Resampling Techniques: Use oversampling, under-sampling, or SMOTE if the target variable is imbalanced.

10. Analyze and Validate Assumptions

- Check for Multicollinearity: Use VIF (Variance Inflation Factor) to detect multicollinearity among predictors.
- Normality Testing: Test if numerical data follows a normal distribution (e.g., using the Shapiro-Wilk test).
- Homoscedasticity: Check the equality of variance across groups.

11. Preliminary Insights and Hypotheses

- Identify Key Findings: Summarize key insights derived from the analysis.
- Generate Hypotheses: Formulate potential hypotheses based on the EDA to be tested in further analysis.

12. Document and Communicate Findings

- Create Visual Summaries: Prepare visualizations that effectively communicate your findings.
- Write a Summary Report: Document the EDA process, key findings, and next steps.
- Presentation: Prepare a presentation if needed, highlighting critical insights and their implications.

13. Next Steps

- Plan for Further Analysis: Decide on the next steps based on your EDA, which may involve modeling, deeper statistical analysis, or further data collection.

14. Review and Reiterate

- Review EDA: Ensure the analysis is comprehensive and address any gaps.
- Iterate as Needed: Depending on findings, revisit earlier steps for further refinement.

This structured approach ensures that the dataset is thoroughly examined and understood before any modeling or deeper analysis begins.