Notes by Izam

Common Steps

- ***Data Collection***
  - *Collect data from various sources*
  - *Scrap data if needed*

- ***Data Cleaning***
  - *Handle Missing*
    - *Imputation (mean, median, mode, regression)*
    - *Remove records with missing*
  - *Handle Duplicate*
    - *Identify and delete duplicate records*
  - *Data transformation*
    - *Change data types*
    - *Standardization (z-score scaling)*
    - *Normalization (min-max scaling)*
    - *Logarithmic transformation*

- ***Data Exploration***
  - *Summary Statistics*
    - *Mean, median, mode*
    - *Variance, standard deviation*
    - *quartiles*
  - *Data Distribution*
    - *Histograms*
    - *Density plots*
    - *Kernel Density Estimation*
  - *Relations*
    - *Scatter plots*
    - *Correlation matrices*
  - *Grouping and aggregation*
    - *Pivot tables*

- ***Data Visualisation***
  - *Use visualization techniques*
    - *Scatter plot*
    - *Histogram*
    - *Box plot*
    - *Bar chart*
    - *Heatmaps*
    - *Line plot*
    - *Violine Plot*
  - *Use Matplotlib, seaborn, and Plotly*

- ***Outlier Detection***

    - *Identify outliers with Univariate, bivariate, and multivariate*
        - *Box plot and whisker plots*
        - *Z-score and IQR*
        - *Visualise like a violin plot or scatter*
    - *Decide whether to treat, remove, or retain outliers*

- ***Feature Engineering***
    - *Create new features based on existing features*
    - *Standardize and normalize data if necessary*
    - *Encode categorical variables*
        - *One hot or label encoding*
    - *Binning or discretization of continuous values*

- ***Hypothesis Testing***
    - *Conduct statistical tests to confirm or reject a hypothesis*
    - *T-test, ANOVA, chi-square*
    - *Correlation*
        - *Pearson*
        - *Spearman*

- ***Dimensionally Reduction***
    - *Use various other techniques*
        - *PCA*
        - *LDA*
        - *t-SNE*

- ***Document Findings***
    - *Tell the insights in the data as a story and make it interesting*

Type of Data

- *Numerical data*
- *Categorical data*
- *Nominal data*
- *Ordinal data*
- *Interval data*
- *Ratio data*
- *Binary data*
- *Sales data*
- *Geographical data*
- *DICOM data*
- *Time series*
- *Text data*
- *Image data*
- *Audio data*

EDA of Text Data

- ***Data collection and Preprocessing***
  - *Scrap the data*
  - *Remove special characters, HTML tags, and other noise.*
  - *Tokenize the data*
  - *Handle issues such as missing and duplicate*

- ***Basic Text Statistics***
  - *Word count, character count, sentence count*
  - *Average word length and sentence length*
  - *Examine the distribution of text length*

- ***Word Frequency Analysis***
  - *Create word cloud or frequency histograms*
  - *Identify the stop words*

- ***Text Visualisation***
  - *Word clouds*
  - *Barcharts and histogram*

- ***N-gram Analysis***
  - *Utilise NLP libraries such as NLTK or Spacy and extract n-grams (bigram, …)*
  - *Plot n-gram frequency distribution*

- ***NER (Name Entity Recognition)***
  - *Use spacy NER models*

- ***Topic modeling***
  - *Apply topic modeling techniques such as Latent Dirichlet Allocation (LDA) or Non-Negative Matrix Factorization (NMF)*

- ***Word Embeddings***
  - *Train or use pre-trained word embedding models like Word2Vec, GloVe, or FastText.*
  - *Visualize word embeddings using techniques like t-SNE*

- ***Language Modeling***
  - *Train and evaluate language models using libraries like TensorFlow, PyTorch, or Hugging Face Transformers*