**REPORT**

PART 1:

1. Method used for Data pre-processing –
   * Understood the overview of the dataset
   * Dropped columns which were irrelevant for prediction, column names -
   * Check for columns with Nan(empty) values or mismatching data types
   * Replace them with mean and mode
   * Check for outliers and removed them
   * Converted Objects to categorical values using one-hot encoding, label encoding and frequency encoding
2. Data Set –
   * Breeding Bird Atlas
     + Domain - Breeding bird observations based on geographical locations
     + Type of Data - The dataset consists of categorical and numerical data
     + Features – 15 features
     + Number of Samples – 361582
     + Mean for each numerical column:
       - Fed. Region – 5.85
       - Month – 49.77
       - Day – 49.5
       - Year – 1964.18
       - Temperature – 49.46
       - Average UB Student – 2.85
     + Standard Deviation for each numerical column:
       - Fed. Region – 5.83
       - Month – 28.65
       - Day – 28.79
       - Year – 190.06
       - Temperature – 17.32
       - Average UB Student – 0.49
     + Missing Values for each column:
       - Fed. Region – 5795
       - Block ID – 2718
       - Map Link – 4717
       - County – 10602
       - Common Name - 10530
       - Scientific Name – 7485
       - NYS Protection Status – 8470
       - Family Name – 2456
       - Family Description -4733
       - Breeding Behavior -5183
       - Month – 358156
       - Day – 352244
       - Year – 10480
     + Graph-
       - Box Plot are used for detecting outliers, for column ‘Year’ , we could see there are few values which lies beyond the first and third quartiles
         * A graph with numbers and a number of objects

           Description automatically generated with medium confidence
         * For Fed. Region we do not have any outliers A blue rectangular object with black lines

           Description automatically generated
         * Bar Plot for County