**Exploratory Data Analysis**

**Datasets:**

**Sports**: ODI batting data, IPL deliveries data

**HR Analytics**: Attrition Data

**Retail**: Super Store data

**Social Media**: Narendra Modi tweets, Flipkart tweets

**Government**: Parliament

**Banking**: Bank campaign data

**Link for data sets:** <https://github.com/skathirmani/datasets>

**Guidelines:**

1. Granularity – Daily/Weekly/Statewise/Citywise/Item level/match level/ball level
2. Data Types
   1. General types: Category, Numerical, Location, Text, Date, Misc
   2. Special types: IDs, Phone numbers, URL, Names
3. Derived Metrics
   1. Date:
      1. Month, Day, Year, Hour, Minute, Second
      2. Afternoon, Morning
      3. Weekday
      4. Week number
      5. Season, age, weekday, weekend
      6. Business hours
      7. Quarter
   2. Location
      1. GDP
      2. Area
      3. Population
      4. Continent
      5. Time zone
      6. Major Language, Total language
      7. Major religion
      8. Ruling party, count of major parties
      9. Currency
      10. Literacy rate
      11. Gender ratio
      12. Unemployment ration
   3. Email ID
      1. Domain
      2. Server
      3. Length of the email address
      4. Special Character (Yes/No)
   4. Name
      1. First Name
      2. Sur Name
      3. Gender
   5. Images
      1. Object
      2. RGB ratio
      3. Theme color
   6. Text
      1. No. of characters, words, sentences
      2. Most repeating word
      3. Noun from the text
      4. Positive, Negative and Neural
      5. No. of hashtags
   7. Numerical columns: Binning
4. Reclassify your data types
5. Identify percentage of missing values
6. Identify percentage of outliers for numerical columns
7. Identify those columns which might not be useful for analysis
8. Univariate Analysis
   1. Category: Frequency distribution to check what percentage of levels contribute to 80% of the frequency
   2. Numerical: Box plot & histogram to identify distribution and outliers
   3. Text
      1. Bag of words - wordcloud
      2. Derive new columns based on top words
   4. Location: Get lat,lon and visualize them using darrinward API
   5. Date: Derive metrics like year, month, day, weekday, quarter etc
9. Bivariate Analysis
   1. Numerical columns – Correlation Analysis
   2. One categorical column & one numerical column – Segmented analysis
      1. T-test or ANNOVA
   3. Two categorical column – Crosstab
      1. Chi-square test
10. Multivariate Analysis
11. Trending Analysis
12. Geographical Analysis
13. Distribution Analysis
14. Relationship Analysis

**Frequency distribution on Categorical Columns**

* Identify unique no. of levels
* For each level identify the frequency
* Sort the levels based on frequency (descending order)
* For each level calculate the %of Freq.
* Calculate cumulative sum of these %of Freq
* Draw the line chart
  + X axis – levels
  + Y axis - % of freq
* Check how many levels contribute to 80% of frequency
* Compute what is the percentage of these levels. If it is less than 25% or so, it is a good pattern to report

**Numerical column: Distribution using box plot**

* Q1, Q2, Q3, IQR = Q3-Q1
* Using IQR
  + Lower whisker -> Q1 – 1.5\*IQR
  + Upper whisker -> Q3 + 1.5 \* IQR
* Using percentiles
  + Lower whisker -> 2 percentile
  + Upper whisker -> 98 percentile
* Identify total number of values which are below the lower whisker and above upper whisker
* Compute % of outliers

**Correlation Analysis**

* Is a measure of linear association between two numerical variables
* The underlying calculation is defined in terms of variances and covariances, in practice we will be using sample correlation coefficient (often denoted as *r*). This is calculated as