**Types of Data**

* **Qualitative**
  + Categorical data – fits into categories.
  + Not numerical.
  + A variance which able to describe quality of the population.
  + **Nominal**
    - labels the variables without providing the numerical value.
    - Nominal scale.
  + **Ordinal**
    - Follows a natural order.
    - A type of data that follows a natural order.
    - Used for visualizations, can be expressed using tables.
* **Quantitative or Numerical**
  + **Discrete**
    - Can take only discrete values.
    - Contains only a finite number of possible values.
  + **Continuous**
    - Can be calculated
    - Infinite number of probable values that can be selected within a specific range.

**7 Types of Analysis**

* Understand the data
* Understand the characteristics

1. **Univariate Analysis**
   1. Univariate analysis is the simplest form of analyzing data.
   2. “Uni” means “one”, so in other words your data has only one variable.
   3. Can be done for numerical and categorical data.
      1. **Numerical** – helps us know the data spread is.
      2. **Categorical** – helps us know the categorical distribution.
   4. Gives us more insights.
   5. Get to know the summary statistics of the particular column.
   6. Does not give any explanation, but it provides an insight already.
2. **Bivariate Analysis**
   1. Data which has two variables, you often want to measure the relationship that exists between these two categorical variables.
   2. Can also be performed with numerical values, or a combination of numerical and categorical values.
   3. Can also be done for numerical and categorical values.
      1. **Numerical**
         1. Covariance & Correlation
            1. Positive
            2. No Correlation
            3. Negative Correlation
            4. Scatter plot / Boxplot
      2. **Categorical**
         1. Quantities / counts per variable.
3. **Multivariate analysis**
   1. Data which has more than two variables, you often want to measure the relationship that exists between these features.
   2. Can be skipped if no more time. Start from uni, bi, and then multi.
4. **Numerical analysis**
   1. For numerical data
   2. Ex. Knowing the single variable:
      1. Mean
      2. Median
      3. 25th percentile
      4. 75th percentile
      5. Min
      6. Max
   3. If more than one
      1. Check their correlation and covariance.
      2. Comparison of box plot, and scatterplot
   4. For several numerical columns
      1. Use corr() function and map it using heatmap.

**Derived Metrics**

* Derived metrics create a new variable from the existing variable to get an insightful information from the data by analyzing the data.
* Variables that we manually create using certain techniques
  + **Feature Encoding**
  + **Feature Binning**
  + **From Domain Knowledge**
  + **Calculated from Data**

**Feature Binning**

* Converts or transform continuous / numeric variable to categorical variable.
* It can also be used to identify values or outliers.
* **Unsupervised Binning**
  + Equal width binning
    - 0-12
    - 13-24
  + Done manually
  + Equal frequency binning (equal datapoints)
    - Split the frequency based on the total number of records automatically.
* **Supervised Binning**
  + Transforms continuous variable into categorical value taking dependent variable into consideration.
  + Entropy based binning
    - **Separate the continuous or numeric variable majority of values in a category belong to same label of class.**
* **Feature Encoding**
  + Feature encoding help us to transform categorical data into numeric data.
  + **Label encoding**
    - Transform categorical variables into numerical variables by assigning a numerical value to each of the categories.
    - Can be biased, never be used on machine learning.
  + **One-hot encoding**
    - This technique is used when independent variables are nominal.
    - It creates k-different columns each for a category and replaces one column with 1 rest of the columns is 0.
    - Here, 0 represents in the absence, and 1 represents the presence of that category.
  + **Target Encoding**
    - In target encoding, we calculate the average of the dependent variable for each category and replace the category variable with the mean value.
  + **Hash Encoder**
    - The Hash encoder represents categorical independent variable using the new dimensions.
    - Here, the user can fix the number of dimensions after transformation using component argument.

**Feature Encoding**

* Help us to transform categorical data into numeric data.

**Label encoding**

* Technique to transform categorical variables into numerical variables by assigning a numerical value to each categories.
* **Can be done on y (column)** variables, but not on x variables.
* Not recommended for record based assignment (ex. Male = 1, Female = 2) because it may cause bias.

**One-hot encoding**

* This technique is used when independent variables are nominal.
* It creates k-different columns each for a category and replaces one column with 1 rest of the columns is 0.
* 1 = True, 0 = False
* Most preferred.
* **Dummy encoding** – removing columns (that are unnecessary)

**Target encoding**

* We calculate the average of the dependent variable for each category and replace the category variables with the mean value.

**Hash Encoder**

* The hash encoder represents the categorical independent variable using the new dimensions.
* Here the user can fix the number of dimensions after transformation using component argument.