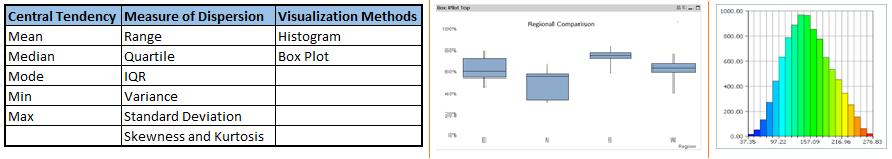
1. **[Steps of Data Exploration and Preparation](https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/" \l "one)**
   * Variable Identification
   * Univariate Analysis



df.boxplot(): by each columns, or specific columns, and by variables

group numerical data through quantiles, Q1-Q3 on box, whiskers is default 1.5 \* (Q3-Q1). Outliers are those past the end of whiskers

histogram:

matplotlib.pyplot as plt

df[‘a’].plot(), .cumsum(), (kind=’bar’) <hist, box, kde, area, scater, pie, hexbin>

or plot.bar(stacked=True) , plot.hist(bins=28), <by=.. by each group>

color=dict(boxes=’Green’, whiskers=’Orange’,medians=’Blue’,caps=’Gary’}

df.plot.box(color=colr, sym=’r+’)

scatter\_matrix(df, alpha=.5,figsize(6,6), diagonal=’kde’)

* + Bi-variate Analysis
    1. **Continuous & Continuous:** scatter plot

**Heatmap:**

**Seaborn.heatmap(df.corr(), cmap=’BuGn’)**

**seaborn.heatmap(corr, xticklables=corr.columns,yticklabels=corr.columns)**

* + 1. **Categorical & Categorical:**
       - **2-way table pd.crosstab()**
       - **stacked column charts**

**dv=df.pivot(index=’yr’, columns=’mth’,values=’Value’)**

**dv.loc[:,[‘jan’,’feb’]].plot.bar(stacked=True,color=, figsize=(6,6))**

**or group**

**df.groupby([‘col1’,’col2’[).size().unsack().plot(kind=’bar’,stacked=True)**

* + - * **CHI-squrare**
    1. Categorical & Continuous: draw box plots for each level of categorical variables. If levels are small in number, it will not show the statistical significance. To look at the statistical significance we can perform Z-test, T-test or ANOVA
       - Z-test or T-test
  + Missing values treatment
  + Outlier treatment
  + Variable transformation
  + Variable creation

1. [**Missing Value Treatment**](https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#two)
   * Why missing value treatment is required ?
   * Why data has missing values?
   * Which are the methods to treat missing value ?
     1. Deletion
     2. Mean/mode/median imputation
2. [**Techniques of Outlier Detection and Treatment**](https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#three)
   * What is an outlier?
   * What are the types of outliers ?
   * What are the causes of outliers ?
   * What is the impact of outliers on dataset ?
   * How to detect outlier ?

Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR

Use capping methods. Any value which out of range of 5th and 95th percentile can be considered as outlier

Data points, three or more standard deviation away from mean are considered outlier

Outlier detection is merely a special case of the examination of data for influential data points and it also depends on the business understanding

Bivariate and multivariate outliers are typically measured using either an index of influence or leverage, or distance. Popular indices such as Mahalanobis’ distance and Cook’s *D* are frequently used to detect outliers.

* + How to remove outlier ?

1. [**The Art of Feature Engineering**](https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#four)
   * What is Feature Engineering ?
   * What is the process of Feature Engineering ?
   * What is Variable Transformation ?
   * When should we use variable transformation ?

Scale change

Change non-linear to linear

Skewed dist to symmetric distribution

Binning of variables (categorization for easy implementation)

* + What are the common methods of variable transformation ?
    1. Logarithm (reduce right skewness )
    2. Square/cube foot (cube can for negative values)
    3. Binning
  + What is feature variable creation and its benefits ?
    1. Derived variable from existing
    2. Dummy variable (categorical to numerical)

<https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/>

**teps in creating competitive advantage while building predictive models:**

There are 2 key steps involved every time you try and create this competitive advantage:

1. Cleansing and transforming datasets to create a comprehensive data set which captures all possible hypothesis
2. Choosing the right set of predictive modeling tools and techniques which work on the data set created in last step and bring out insights

**How do we extract maximum information from a defined data set?**riable

1. Create variables for difference in date, time and addressesfor difference in date, time and addresses

2. Create new ratios and proportions

3. Apply standard transformations (log, exponential, quadratic, trignomtric)

4. Include effect of influencer (like staff size of org.)

5. Check variables for seasonality and create the model for right period

**List of Common Machine Learning Algorithms**

Here is the list of commonly used machine learning algorithms. These algorithms can be applied to almost any data problem:

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. kNN
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms
10. Gradient Boosting algorithms
    1. GBM
    2. XGBoost
    3. LightGBM
    4. CatBoost

<https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>

Dimension Reduction Techniques

<https://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/>

Table of Contents

1. Why Dimension Reduction is Important in machine learning and predictive modeling?
2. What are Dimension Reduction techniques?
3. What are the benefits of using Dimension Reduction techniques?
4. What are the common methods to reduce number of Dimensions?
   1. Missing values: impute or drop variables (40%+ missing if)
   2. drop variables having low variance compared to others because these variables will not explain the variation in target variables.
   3. Decision tree
   4. Random forest
   5. High correlation (pearson/ polychoric) correlation matrx
   6. Backward feature elimination
   7. Factor analysis
   8. PCA (principle component analysis)
5. #Import Library
6. from sklearn import decomposition
7. #Assumed you have training and test data set as train and test
8. # Create PCA obeject pca= decomposition.PCA(n\_components=k) #default value of k =min(n\_sample, n\_features)
9. # For Factor analysis
10. #fa= decomposition.FactorAnalysis()
11. # Reduced the dimension of training dataset using PCA
12. train\_reduced = pca.fit\_transform(train)
13. #Reduced the dimension of test dataset
14. test\_reduced = pca.transform(test)
15. Is Dimensionality Reduction good or bad?