# SI 618

## Pandas DataFrame

Creating dataframes

```
df = pd.DataFrame({'a':[4,5,6], 'b':[7,8,9], index=[1,2,3]})
```

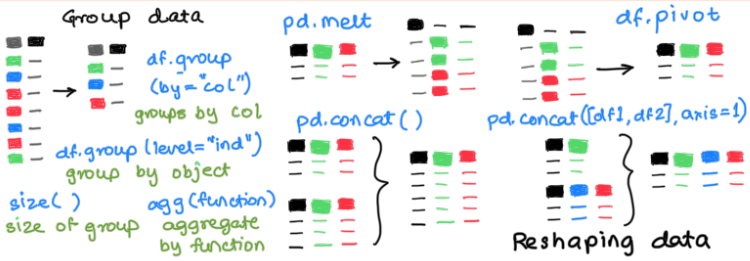
**Read**
- `df = pd.read_csv("path")` read CSV file to DF
- `pd.read_excel("path")` read excel file to DF

**Basic**
- `df.head` first 5 rows
- `df.tail` last 5 rows
- `df.shape` dimensions
- `df.info` information
- `df.describe` summary
- `df.index` index info

**Summary**
- `sum()` `min()`
- `count()` `max()`
- `median()` `mean()`
- used to summarize

**Missing**
- `df.dropna()` Drop rows with any column having N/A
- `df.fillna()` Replace N/A with value

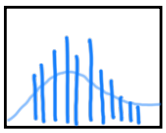
**Select**
- Extracting Rows `df.loc[0]` `df.iloc[0]`
- Select element by position `df.iloc[0,1]`



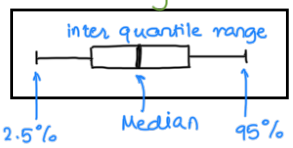

## Visualization by Seaborn

### Univariate Analysis

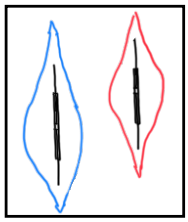
`sns.distplot` gives distribution



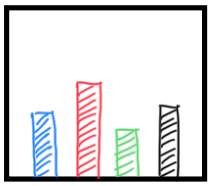
`sns.boxplot` distribution of quantitative data on summary statistics




`sns.violinplot` presents data distribution through density and quantile information

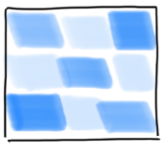


`sns.countplot()` used to show `value_counts()` of each feature of data

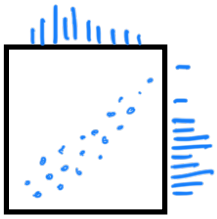


### Bivariate Analysis

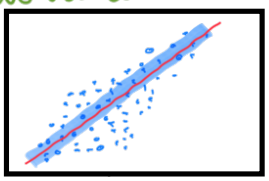
`heatmap()` displays color encoded matrix for correlations



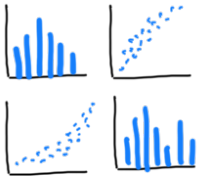
`jointplot()` bivariate data with scatter and univariate histograms



`regplot()` used to depict the regression fit b/w two variables and CI



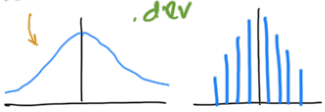
`pairplot()` creates a grid of pairwise relationships between variables in a dataset



## Statistics

### Types of Distributions

Normal (Gaussian) Bell shaped curve with mean and std. dev



Binomial distr. # successes in fixed trials

Continuous

Uniform all values have equal probability

Discrete

Poisson count of events occus at equal time intervals

### Test Statistic

Z-test - compares sample mean to a known population mean when pop. std. dev known

T-test - determines if there is significant difference b/w means of two groups

$X^2$ - test - Asses independance b/w two categorical variables

ANOVA (Analysis of variance) compares means of more than two groups ⇒ significant

### Hypothesis Testing

$H_0$ (Null Hypothesis) the two data are similar

$H_A$ (Alternate Hyp) the data are dissimilar or thus regression is significant

significance level ($\alpha$) - typically 0.05 is considered critical value. ⇒ confidence interval

## Categorical Data

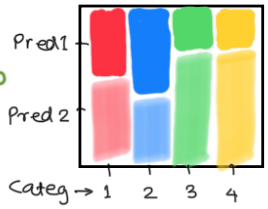
Contingency Tables displays frequency & counts of categorical features

Cross Tab summarizes and computes frequencies of variables

Chi-Squared Test Tables easy to compare b/w observed and expected values

Mozaic plot gives visual interpretations

use combination of visualization and statistic-testing




## Text Data

- `.lower()` converts to lower case
- `.upper()` converts to upper case
- `.split()` divides string based on a delimiter
- `.replace()` used to replace characters, strings or patterns

### Regular Expressions

`'\d{3}-\d{3}-\d{4}'` Matches phone: xxx-xxx-xxxx

`'[A-Za-z]+'` Matches one or more alphabetical characters

`'^\d{4}$'` Matches a string that contains exact 4 digits

They are powerful tools for string matching, pattern extraction and manipulation in text data.

These form a part of preprocessing in NLP tasks.

Tokenization - breaking text into words/phrases

Cleaning - Removal of unnecessary characters, symbols or white-spaces

Normalization - standardizing text by converting to lower case, removing punctuation etc. sometimes extracting root word / "lemmatization"

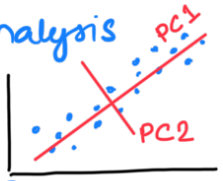
# WHAT DO YOU WANT TO DO ?

## DIMENSION REDUCTION

Reduce data dimension

TECHNIQUES

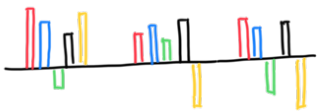
- **Principle component Analysis**
  Projects data into orthogonal components to max. variance
  choice of number of P.C.




- **Multi Dimensional Scaling**
  Reduces dimension while keeping pairwise distance
  linear and non linear data
- **T-SNE**
  emphasizes local and global structure preservation in high dimensions
  tune in perplexity for good results

EVALUATION

- Explained variance by each component



## CLUSTERING

Discover the structure

### K-Means

divides data into 'K' clusters based on centroids
better for spherical

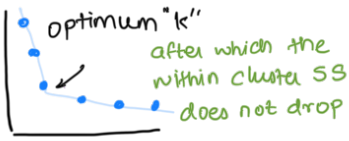
### Heirarchical

Forms agglomerative clustering based on proximity
better when # clusters is not known

TECHNIQUES

EVALUATIONS

### Elbow Method



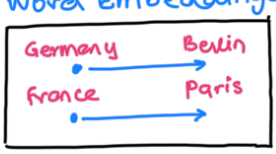

### Silhouette Method

silhouette score [-1, 1] shows how well a data point is clustered


## NATURAL LANGUAGE PROCESSING

Extract Information from the text

- Preprocess the text: punctuation, STOP_WORDS,
- Vectorize the text:
  Bag of Words : one Hot encoded dictionary
  TFIDF : keeping relevant words in matrix, utility
- POS tagging - labelling words with their POS
  - Named Entity Recognition named entities like locations, names, companies, events
- Word Embeddings → Represent each word as real valued vectors




→ Cosine Similarity to identify similar words using NLTK

- Sentiment Analysis

## MACHINE LEARNING

Predict the outcome

scikit-learn is machine learning library

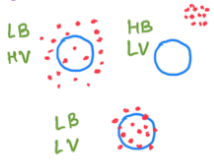
CONCEPTS

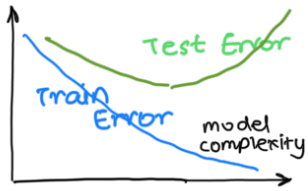
PROCEDURE

Data Preprocessing
- train, validation test split
- scaling, encoding categorical

Import the model
Choose hyperparameters
Arrange data: features, labels
fit() your model
predict() on new data

### Bias-Variance



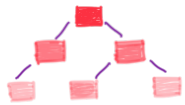

MODELS

**Regression** continuous
- interpretability ← Linear Regression()
- complex relations ← Random Forest Regressor()

**Classification** Discrete Classes
- binary classifier ← Logistic Regressor()
- non-linear kernels ← support Vector Machines()
- categorical variable ← Decision Tree classifier()

### Feature Importance

represents which feature has higher importance



feature at higher split are more important

### Cross Validation

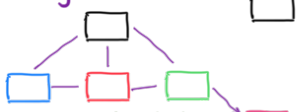
is a resampling technique that uses different portions of data to test performance by hyper-param

TECHNIQUES

### Ensemble techniques

Bagging - combine several weak learners

Boosting



build a strong learner by learning from errors of previous classifiers.

Voting Classifier takes average of model outputs

PIPELINES helps to group together data preprocessing steps

Stacking classifier allows to use strength of each classifier

### Network Architecture :