- Heatmap
  boxplot
  countplot
  barplot

Pearson's correlation

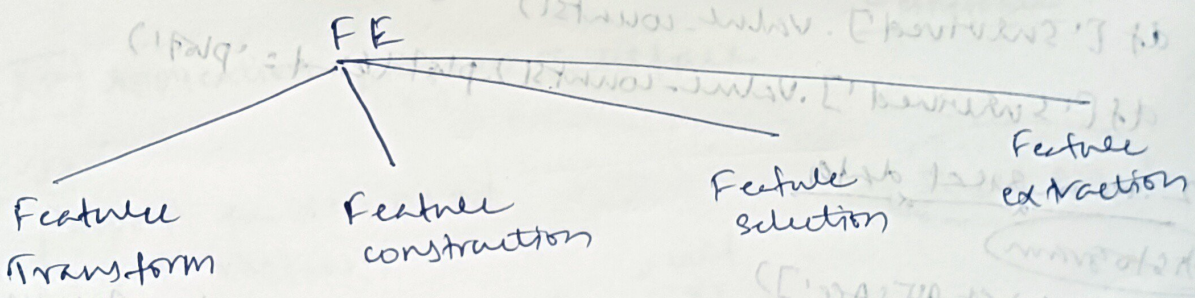
Pandas Profiling

```
from pandas_profiling import ProfileReport
Prof = ProfileReport(df)
prof.to_file(output_file="output.html")
```

- makes a webpage which contains the report of our data.

Feature Engineering

- The process of using domain knowledge to extract features from raw data.



- Missing value imputation
- Handling categorical features (Encoding)
- Outlier detection
- Feature scaling

Standardization

- also called z-score normalization

Age	salary
27	
15	
33	
63	
:	
500 rows	

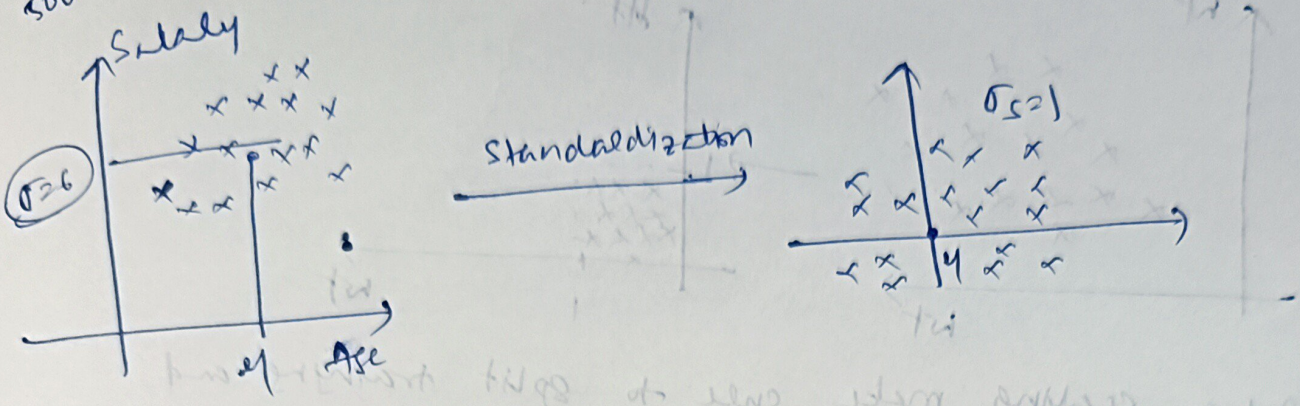
$$x_i' = \frac{x_i - \bar{x}}{\sigma}$$

$\mu = 0 \quad \sigma = 1$

after standardization.

age | salary

500



Standard scaler

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(X_train)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

scaler.mean_

- scaling doesn't effect decision tree but effects logistic regression.

- K-means
- KNN
- PCA
- ANN
- Gradient descent

} use standardization.

Normalization

- To change the values of numeric columns in the dataset to use common scale.

- MinMax scaling
- Robost scaling
- Mean Normalization
- Max absolute

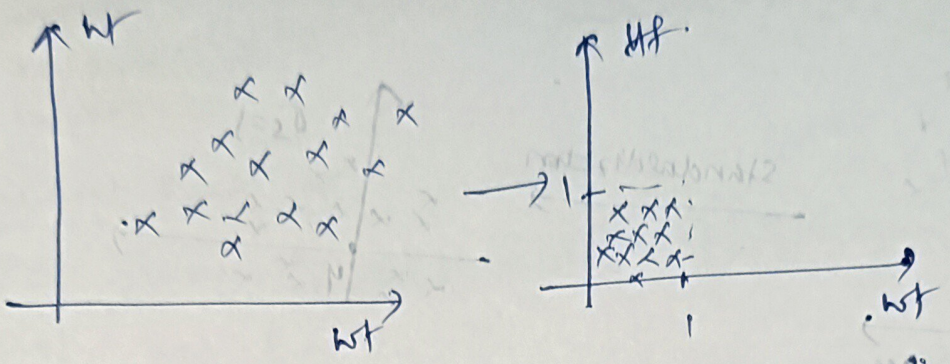
- sets the data in a single range

MinMax Scaling

Weights
130
67
92
81

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Wt, Ht



- Before scaling make sure to split training and testing data.

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaler.fit(X_train, Y_train)
X_scaled_train = scaler.transform(X_train)
X_scaled_test = scaler.transform(X_test)
```

Mean Normalization: $x_i' = \frac{x_i - x_{mean}}{x_{max} - x_{min}}$

Max Absolute scaling: $x_i' = \frac{x_i}{|x_{max}|}$ (for sparse matrix)

Robust scaling: $x_i' = \frac{x_i - x_{median}}{IQR}$ {75th percentile − 25th percentile}

↓

Robust to outliers