Retrospective analytics vs predictive analytics vs real time analytics (it is a type of predictive analytics but often combined with some sort of decision)

Clustering, classification (predicts a discrete categorical/set/lookup value), regression (predicts a continuous numeric value).

Data in form of Label and features (where label can be class label or clustering label or any other predicted value) as a single row. Phases in general are learning (training or model building), test and evaluate, and at last prediction phase (in production)

For classification: check whether the label values in available dataset (for training and evaluation) is balanced using histogram.

Missing values: either remove rows or substitute a particular value or interpolate values or forward fill or backward fill or impute. Pandas.DataFrame.isnull() safer than Pandas.DataFrame.isna()

Repeated values: Pandas.DataFrame.drop\_duplicates()

Outliers:

Scaling: If some numeric columns have high magnitude values, they might tend to dominate the learning process. So look to scale data. Scikit-learn.preprocessing.Scale(). So scaling is actually normalization?