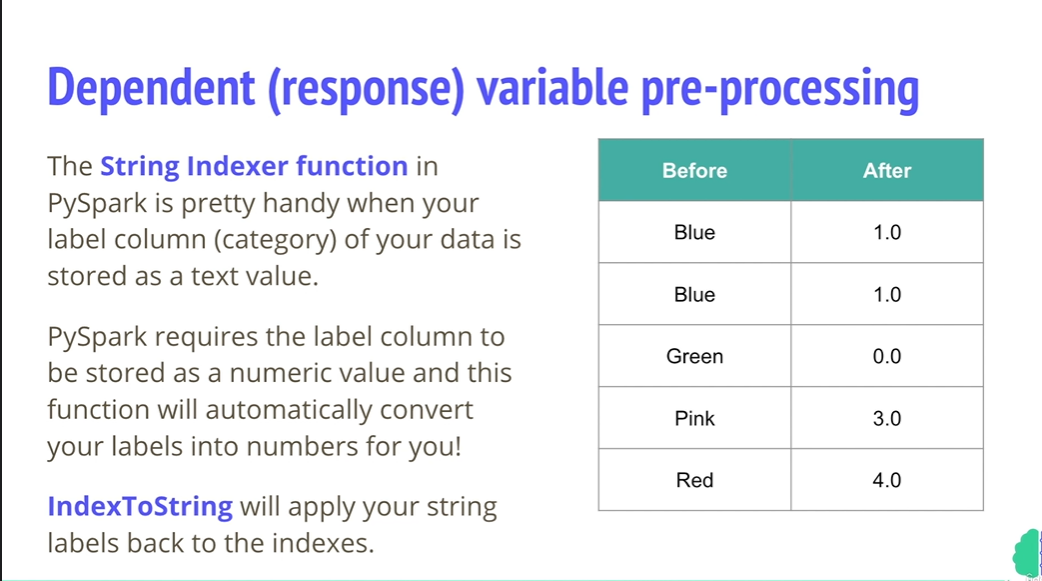
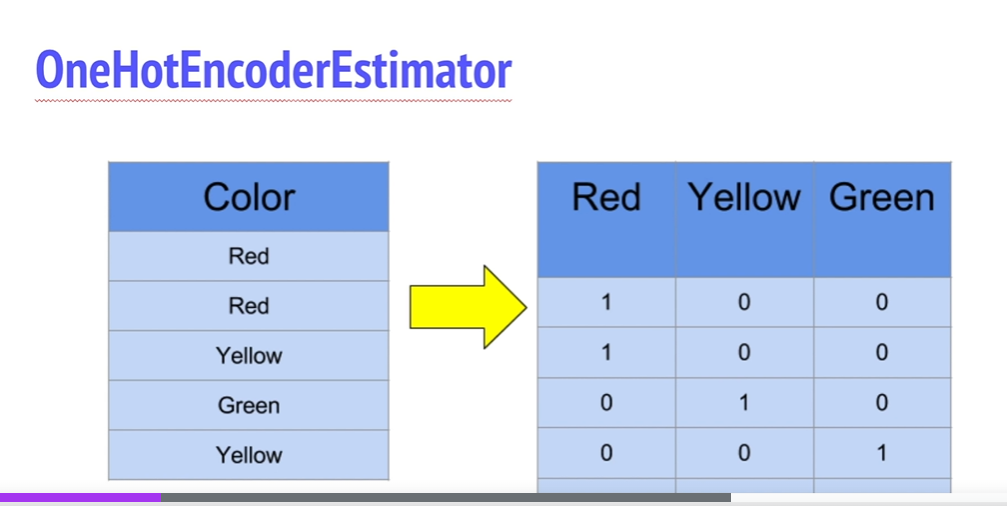
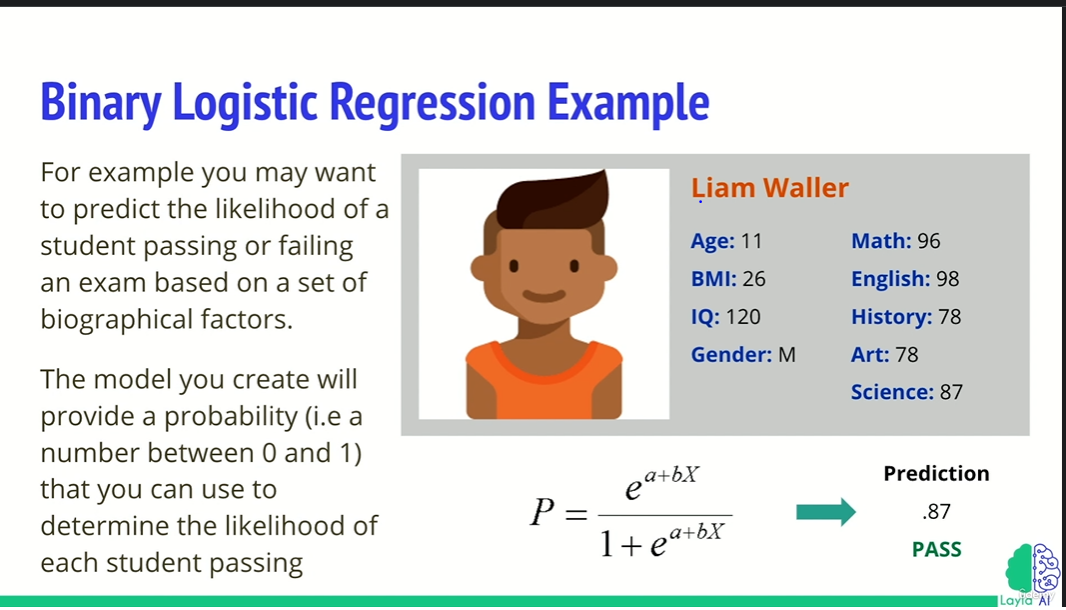
Binary classification:

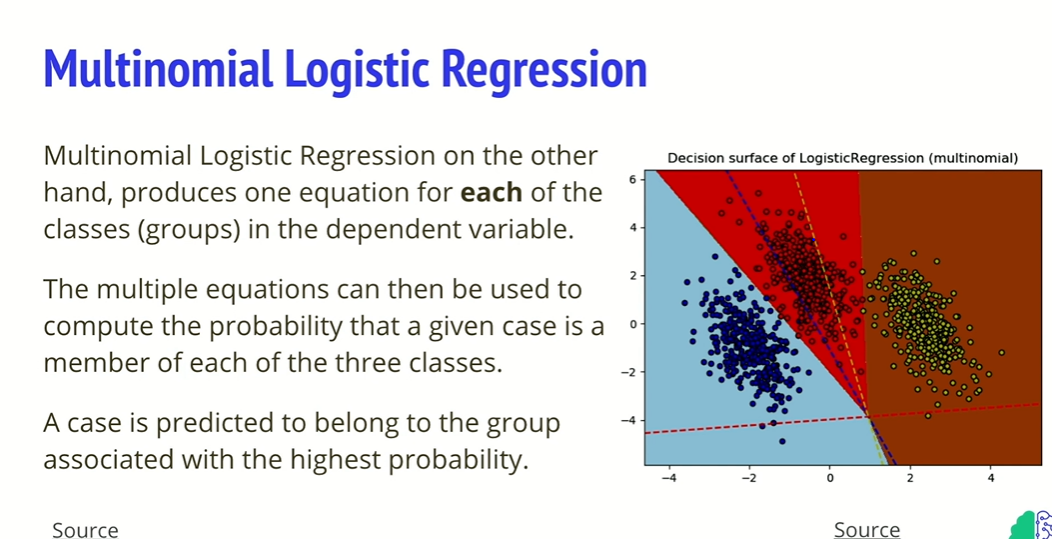


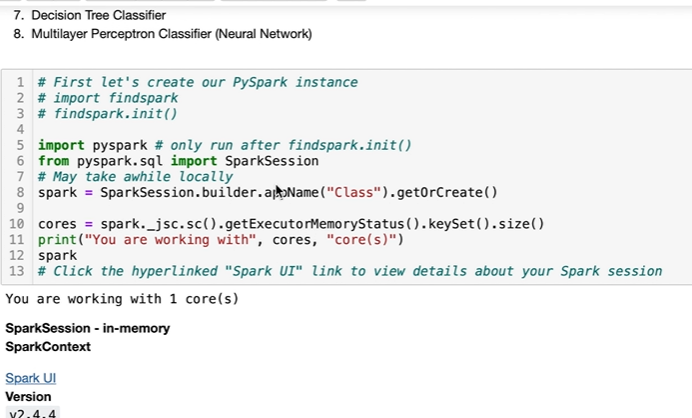
Categorical

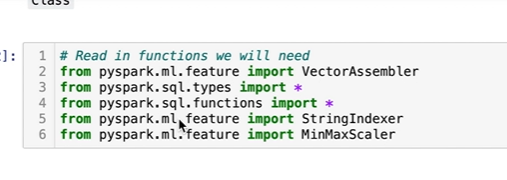
Another method to deal with categorical

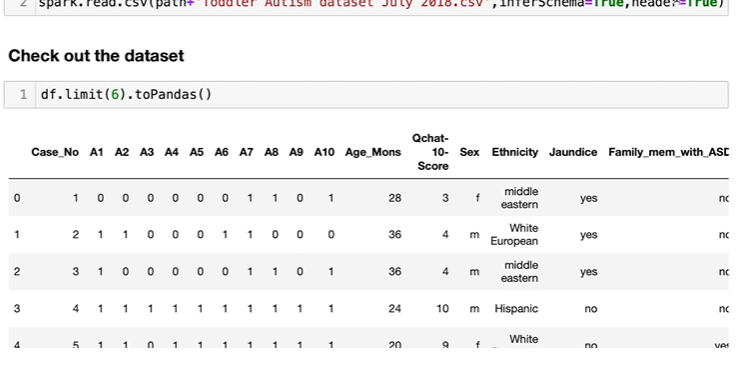




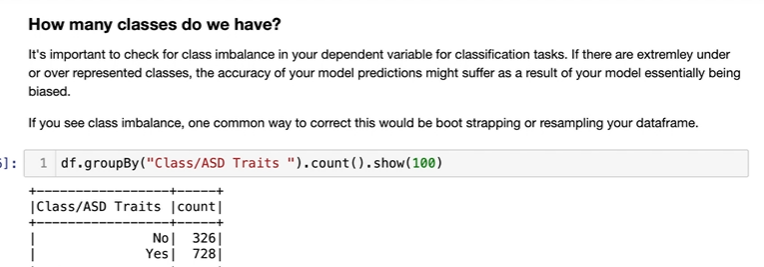


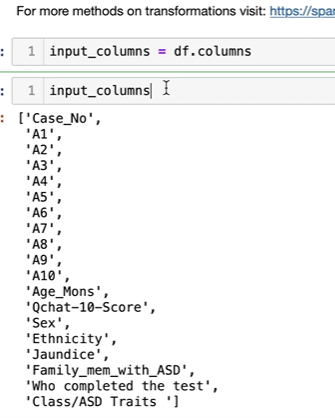


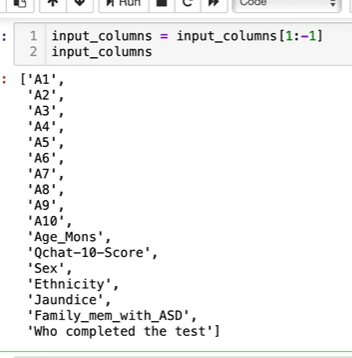


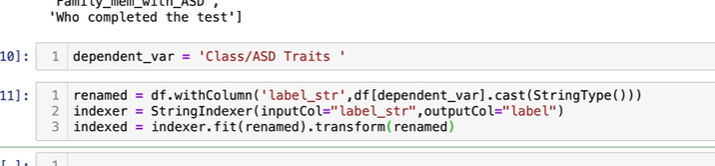


df.printSchema()

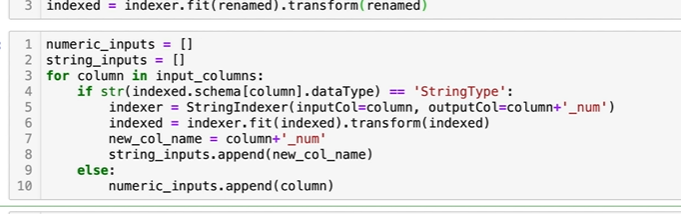








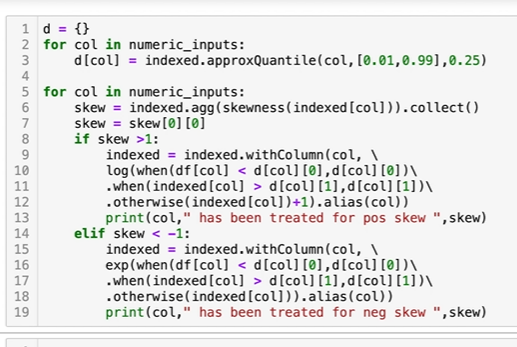
Now converting all string type to string indexer type



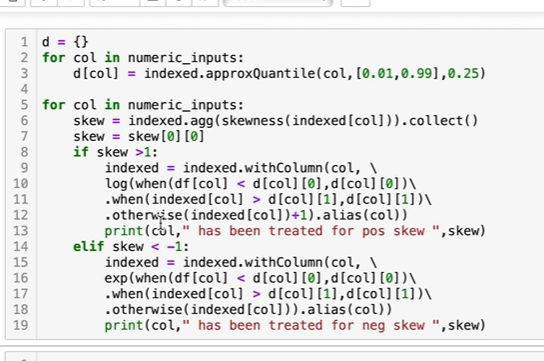
Now dealing with numerical values:

Getting ouliers

We need 99per and 1 percentile



Treating skew for numerical values



Check for negative values in dataframe

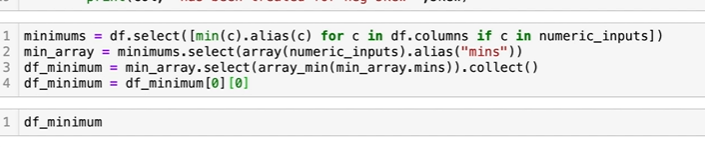
<https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-pyspark-rdd-operations/>

<https://www.mlflow.org/docs/latest/python_api/mlflow.xgboost.html>

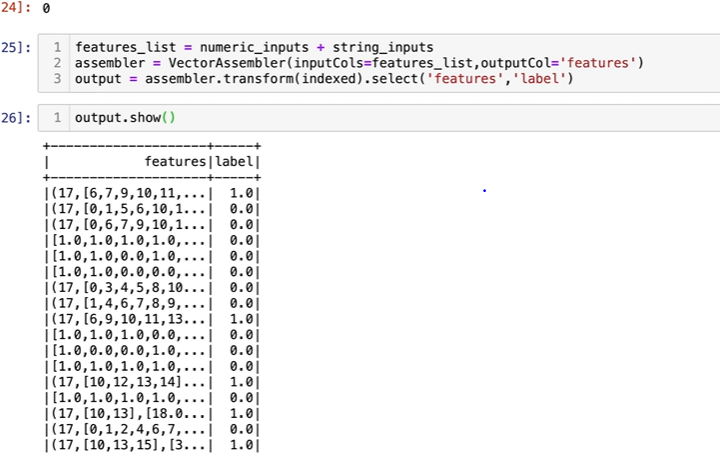
<https://docs.databricks.com/_static/notebooks/mlflow/mlflow-end-to-end-example.html>

finding min value in each column

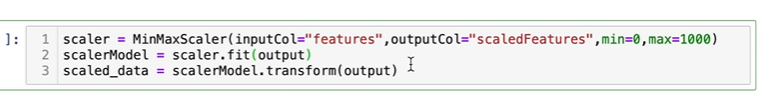
min value in a df

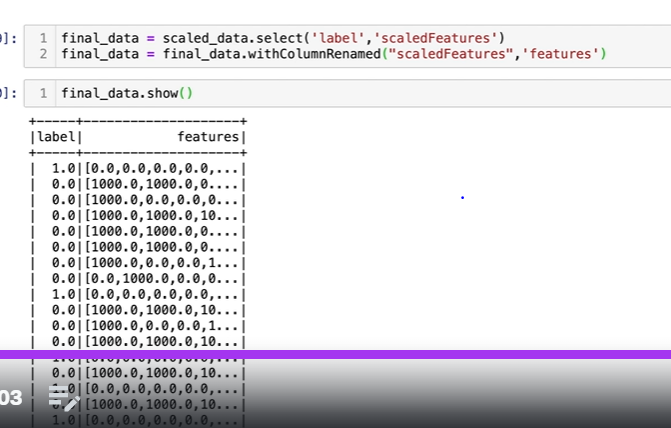


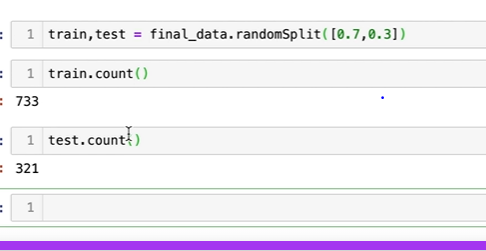
Vector assembler

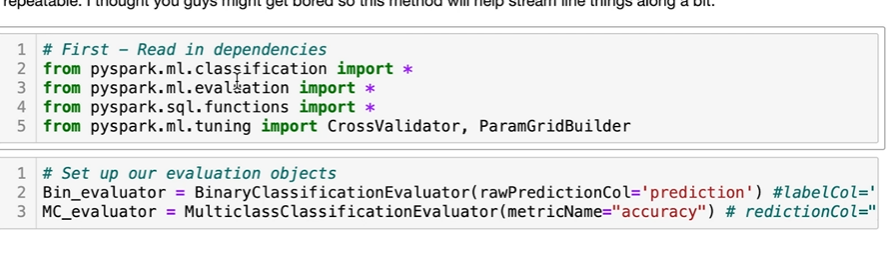


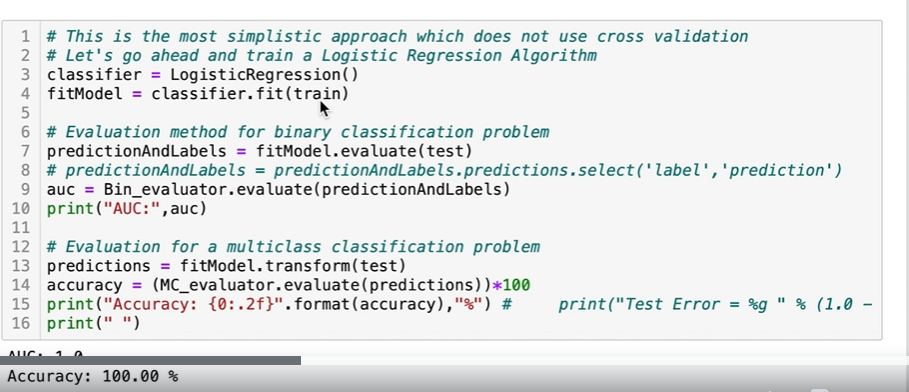
Scaling











Cross validation , maxiters

