Topic	pandas	spark
Read csv file into dataframe	pd.read_csv	spark.read.csv
Display column names	df.columns	df.show
Filter dataframe	df[df["column name"] == condition]	df.filter(df.column name == condition)
Find dimension of dataframe	df.shape	df.count(),len(df.columns)
Group By - Aggregation	df.groupby("column 1").agg({'Column 2':'mean', 'Column 3':'sum'}).reset_index()	df.groupBy("Column 1").agg({'Column 2':'avg', 'Column 3':'sum'})
Rename Columns	df.rename(columns = {"Column 1":"Column 1 renamed", "Column 2":"Column 2 renamed"})	df.withColumnRenamed("Column 1","Column 1 renamed") .withColumnRenamed("Column 2"," Column 2 renamed")
Join	pd.merge(df1,df2,on = ['Column 1'],how = "inner/left/right")	df1.join(df2,['Column 1'],inner/left/right)
Change datatype	df['Column 1'].astype('datatype')	df.withColumn("Column 1",col("Column 1").cast ("datatype"))
Create derived column	df['New Column'] = round((df['Column 1']/df ['Column 2'])*100,2)	df.withColumn("New Column", round((col("Column 1")/col("Column 2"))*100,2))
Sort	df.sort_values(['Column 1'], ascending=[0])	df.sort(df.Column 1.desc())