```
from pyspark.sql.functions import approx_count_distinct,collect_list from pyspark.sql.functions import collect_set,sum,avg,max,countDistinct,count from pyspark.sql.functions import first, last, min, mean from pyspark.sql.functions import sumDistinct simpleData = [("James", "Sales", 3000), ("Michael", "Sales", 4600), ("Robert", "Sales", 4100), ("Maria", "Finance", 3000), ("James", "Sales", 3000), ("Scott", "Finance", 3300), ("Jen", "Finance", 3900), ("Jeff", "Marketing", 3000), ("Kumar", "Marketing", 2000), ("Saif", "Sales", 4100) ] schema = ["employee_name", "department", "salary"] df = spark.createDataFrame(data=simpleData, schema = schema) df.printSchema() df.show() df.select(approx_count_distinct("salary")).show() df.select(avg("salary")).show() df.select(collect_list("salary")).show() df.select(collect_set("salary")).show() df.select(countDistinct("salary").show() df.select(count("salary")).show() df.select(first("salary")).show() df.select(last("salary")).show() df.select(max("salary")).show() df.select(min("salary")).show() df.select(mean("salary")).show() df.select(sum("salary")).show() df.select(sumDistinct("salary")).show()
```