TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE TIJUANA

SUBDIRECCIÓN ACADÉMICA
DEPARTAMENTO DE SISTEMAS Y COMPUTACIÓN

SEMESTRE:
Enero - Junio 2020

CARRERA:
Ing. Tecnologías de la Información y Comunicaciones

MATERIA:
Datos Masivos

UNIDAD A EVALUAR:
Unidad 1

NOMBRE Y NÚMERO DE CONTROL DEL ALUMNO:
Diaz Martinez Ruben Emilio # 15210791

NOMBRE DEL MAESTRO (A):
Dr. Jose Christian Romero Hernandez

28 de Febrero del 2020

```
import org.apache.spark.sql.SparkSession

val spark = SparkSession.builder().getOrCreate()

val df = spark.read.option("header",
"true").option("inferSchema","true")csv("CitiGroup2006_2008")



//1.sumDistinct

df.select(sumDistinct("Sales")).show()



//2.last

df.select(last("Company")).show() //last data in company



//3.first

df.select(first("Person")).show() first data in person



//4.var_pop

df.select(var_pop("Sales")).show()



//5.avg

df.select(avg("Sales")).show()



//6.collect_list

df.select(collect_list("Sales")).show()



//7.var_samp

df.select(var_samp("Sales")).show()
```

```
//8.sum

df.select(sum("Sales")).show()



//9.stddev_pop

df.select(stddev_pop("Sales")).show()



//10.skewness

df.select(skewness("Sales")).show()



//11.min

df.select(min("Sales")).show()



//12.kurtosis

df.select(kurtosis("Sales")).show()



//13.collect_set

df.select(collect_set("Sales")).show()



//14.approx_count_distinct

df.select(approx_count_distinct("Company")).show()



//15.mean

df.select(mean("Sales")).show()



//16 return the first column of the dataframe
```

```
df.first



//17 Returns the dataframe columns

df.columns



//18 Add a column that derives from the high and Volume column

val df2 = df.withColumn("HV Ratio", df("High")+df("Volume"))



//19 Choose the volume column min

df.select(min("Volume")).show()



//20 Choose the volume column max

df.select(max("Volume")).show()
```