UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

MAESTRÍA EN GESTIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN



Big Data

Alumnos: Ledezma Ibarra Juan Francisco López Ramírez Flavio Hiram

Evaluación 1

Octubre 2023, Tijuana Baja California, México



Imagen 1 - Archivo Readme en Github

```
README.md X = evaluationPractice1.scala
                                            ■ df.scala •
                                                           ≡ test.scala
ClaseBigData > 🕦 README.md > 🗪 # Evaluation Unit 1
405
      ## 1.- Comienza una simple sesión Spark.
      ```scala
 import org.apache.spark.sql.SparkSession
 val spark = SparkSession.builder().getOrCreate()
      ```sh
      val args: Array[String] = Array()
     Loading Practices\evaluationPractice1.scala...
      import org.apache.spark.sql.SparkSession
      ## 2.- Cargue el archivo Netflix Stock CSV en dataframe llamado df, haga que Spark infiera los tipos de datos.
      ```scala
 val df = spark.read.option("header", "true").option("inferSchema", "true")csv("Practices/Netflix_2011_2016.csv")
 val df: org.apache.spark.sql.DataFrame = [Date: date, Open: double ... 5 more fields]
```

Imagen 2 - Archivo Fuente de Readme (Visual Code)

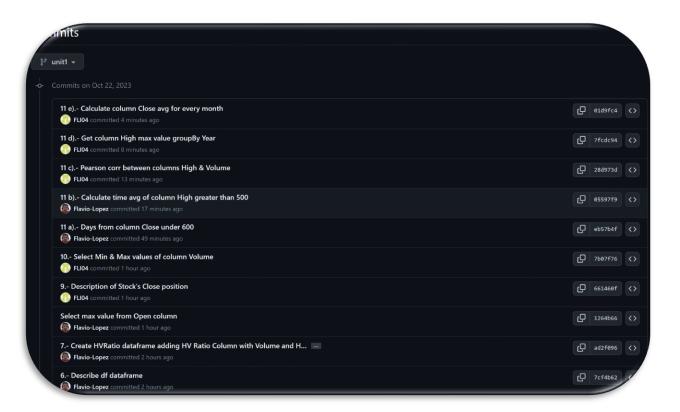


Imagen 3 - Historial de Commits Generados en Github

```
■ df.scala ● ■ test.scala
 ■ evaluationPractice1.scala ×
SigData > Practices > ■ evaluationPractice1.scala
 import org.apache.spark.sql.SparkSession
 val spark = SparkSession.builder().getOrCreate()
 // 2. Cargue el archivo Netflix Stock CSV en dataframe llamado df, haga que Spark infiera los tipos de datos.
val df = spark.read.option("header", "true").option("inferSchema", "true")csv("Practices/Netflix_2011_2016.csv")
 df.printSchema()
 df.printSchema()
 df.head(5)
 df.describe().show()
 val HVRatio = df.withColumn("HV Ratio", $"Volume" / $"High")
 HVRatio.show()
 df.select(max("OPEN")).show()
 df.select(max("Volume"),min("Volume")).show()
 df.filter($"Close"<600).count()</pre>
 (df.filter($"High">500).count()*1.0/df.count())*100
 c) ¿Cuál es la correlación de Pearson entre columna "High" y la columna "Volumen"?
```

Imagen 4 - Código en Scala