Proyecto programación week14

```
%spark

import org.apache.spark.sql.types.StringType
import org.apache.spark.sql.functions._
spark.conf.set("spark.sql.legacy.allowUntypedScalaUDF", "true")

val data = spark
.read
.option("inferSchema", "true")
.option("header", "true")
.option("delimiter", ",")
.csv("/home/carlos/Descargas/PhiUSIIL_Phishing_URL_Dataset.csv")
```

SPARK JOB FINISHED

```
import org.apache.spark.sql.types.StringType
import org.apache.spark.sql.functions._
data: org.apache.spark.sql.DataFrame = [FILENAME: string, URL: string ... 54 more fields]
```

Took 42 sec. Last updated by anonymous at July 22 2025, 7:58:04 PM.

```
%spark
data.printSchema()
```

FINISHED

```
root
 |-- FILENAME: string (nullable = true)
 |-- URL: string (nullable = true)
 |-- URLLength: integer (nullable = true)
 |-- Domain: string (nullable = true)
 |-- DomainLength: integer (nullable = true)
 |-- IsDomainIP: integer (nullable = true)
 |-- TLD: string (nullable = true)
 |-- URLSimilarityIndex: double (nullable = true)
 |-- CharContinuationRate: double (nullable = true)
 |-- TLDLegitimateProb: double (nullable = true)
 |-- URLCharProb: double (nullable = true)
 |-- TLDLength: integer (nullable = true)
 |-- NoOfSubDomain: integer (nullable = true)
 |-- HasObfuscation: integer (nullable = true)
 |-- NoOfObfuscatedChar: integer (nullable = true)
```

```
|-- ObfuscationRatio: double (nullable = true)
```

Took 0 sec. Last updated by anonymous at July 13 2025, 8:26:48 PM.

Proyecto programación week14

FINISHED

```
%spark

def getSecurityCategory(
    IsHTTPS: Int, HasTitle: Int, Robots: Int, HasCopyrightInfo: Int, HasObfuscation: Int, NoOfURLRedirect: Int,
    NoOfPopup: Int, NoOfiFrame: Int, HasExternalFormSubmit: Int, HasPasswordField: Int, Bank: Int, Pay: Int, Crypto: Int,
    URLCharProb: Double, URLLength: Int, NoOfSubDomain: Int
): String = {

    var score = 0

    // Características que suman puntos a favor
    if (IsHTTPS == 1) score += 3
    if (HasTitle == 1) score += 2
    if (Robots == 1) score += 1
    if (HasCopyrightInfo == 1) score += 3

    // Características que restan puntos a favor
    if (HasObfuscation == 1) score -= 2
    if (NoOfURLRedirect > 0) score -= 2
    if (NoOfPopup > 0) score -= 3
    if (NoOfiFrame > 0) score -= 1
    if (HasExternalFormSubmit == 1) score -= 2
    if (HasPasswordField == 0 && (Bank == 1 || Pay == 1 || Crypto == 1)) score -= 4 // Si no necesita password, pero si banco, pago o cripto, r
    if (URLCharProb < 0.5) score -= 2

    // COnsideraciones mínimas que restan puntos a favor.
    if (URLLength > 100) score -= 1
    if (NoOfSubDomain > 3) score -= 1

    // Definir los rangos y retornar la categoría
    score match {
        case s if s >= 5 => "HIGH"
        case s if s >= 0 && s < 5 => "MEDIUM"
        case s if s >= -5 && s < 0 => "LOW"
        case s if s < -5 => "VERY LOW"
        case _ => "UNDEFINED"
    }
}

// Crear función UDF para que pueda ser usada en cada iteración:
val getSecurityCategoryUdf = udf(getSecurityCategory(
    _: Int, _: Int, _: Int, _: Int, _: Int, _: Int, _: Int, _: Int,
    _: Int, _: Int, _: Int, _: Int, _: Int, _: Double, _: Int, _: Int
), StringType)
```

```
warning: there was one deprecation warning (since 3.0.0); for details, enable `:setting -deprecation' or `:replay -deprecation'
getSecurityCategory: (IsHTTPS: Int, HasTitle: Int, Robots: Int, HasCopyrightInfo: Int, HasObfuscation: Int, NoOfURLRedirect: Int, NoOfPopup: In
t, NoOfiFrame: Int, HasExternalFormSubmit: Int, HasPasswordField: Int, Bank: Int, Pay: Int, Crypto: Int, URLCharProb: Double, URLLength: Int, N
oOfSubDomain: Int)String
getSecurityCategoryUDF: org.apache.spark.sql.expressions.UserDefinedFunction = SparkUserDefinedFunction($Lambda$3492/2010635149@49e269b5,String
Type,List(),None,None,true,true)
```

Proyecto programación week1

Took 2 sec. Last updated by anonymous at July 21 2025, 10:59:06 AM.

FINISHED

```
%spark

def getUrlComplexity(NoOfSubDomain : Int, URLLength : Int, NoOfOtherSpecialCharsInURL : Int, SpacialCharRatioInURL : Double) : String = {
    val complexityScore = (NoOfSubDomain * 1.5) + (URLLength * 0.1) + (NoOfOtherSpecialCharsInURL * 1.0) + (SpacialCharRatioInURL * 5.0)
     complexityScore match {
        case score if score < 10 => "SIMPLE"
        case score if score < 30 => "MEDIUM"
        case score if score < 60 => "COMPLEX"
        case score if score >= 60 => "VERY COMPLEX"
        case _ => "UNDEFINED"
    }

}

//Función UDF para que pueda ser usada en cada iteración

val getUrlComplexityUdf = udf(getUrlComplexity(
    _: Int, _: Int, _: Int, _: Double), StringType)
```

```
warning: there was one deprecation warning (since 3.0.0); for details, enable `:setting -deprecation' or `:replay -deprecation'
getUrlComplexity: (NoOfSubDomain: Int, URLLength: Int, NoOfOtherSpecialCharsInURL: Int, SpacialCharRatioInURL: Double)String
getUrlComplexityUdf: org.apache.spark.sql.expressions.UserDefinedFunction = SparkUserDefinedFunction($Lambda$3533/517623571@3a95305a,StringTyp
e,List(),None,None,true,true)
```

Took 1 sec. Last updated by anonymous at July 21 2025, 10:59:10 AM.

FINISHED

```
%spark


def getPresenceRatio(NoOfImage : Int, NoOfCSS : Int, NoOfJS : Int, LineOfCode : Int) : String = {

val totalContentElements = NoOfImage + NoOfCSS + NoOfJS
```

Proyecto programación week1

```
warning: there was one deprecation warning (since 3.0.0); for details, enable `:setting -deprecation' or `:replay -deprecation'
getPresenceRatio: (NoOfImage: Int, NoOfCSS: Int, NoOfJS: Int, LineOfCode: Int)String
getPresenceRatioUdf: org.apache.spark.sql.expressions.UserDefinedFunction = SparkUserDefinedFunction($Lambda$3535/59657740@35ce452,StringType,List(),None,None,true,true)
```

Took 1 sec. Last updated by anonymous at July 21 2025, 10:59:13 AM.

SPARK JOB (http://10.0.2.15:4040/jobs/job?id=2) FINISHED

```
%spark
/*
Poblar la tabla con SecurityScore
*/
val dataWithSecurity = data.withColumn(
    "SecurityCategory",
    getSecurityCategoryUdf(
        col("IsHTTPS"),
        col("HasTitle"),
        col("Robots"),
        col("HasCopyrightInfo"),
        col("HasObfuscation"),
        col("NoOfURLRedirect"),
        col("NoOfPopup"),
        col("NoOfiFrame"),
        col("HasExternalFormSubmit"),
        col("HasPasswordField"),
        col("Bank"),
        col("Pay"),
        col("Crypto"),
```

Proyecto programación week1

```
+---------------------------------+----------------+
|URL                              |SecurityCategory|
+---------------------------------+----------------+
|https://www.southbankmosaics.com |MEDIUM          |
|https://www.uni-mainz.de         |HIGH            |
|https://www.voicefmradio.co.uk   |HIGH            |
|https://www.sfnmjournal.com      |LOW             |
|https://www.rewildingargentina.org|MEDIUM          |
|https://www.globalreporting.org  |MEDIUM          |
|https://www.saffronart.com       |MEDIUM          |
|https://www.nerdscandy.com       |HIGH            |
|https://www.hyderabadonline.in   |HIGH            |
|https://www.aap.org              |HIGH            |
|https://www.religionenlibertad.com|MEDIUM          |
|http://www.teramill.com          |LOW             |
|https://www.socialpolicy.org     |MEDIUM          |
|https://www.aoh61.com            |MEDIUM          |
```

Took 2 sec. Last updated by anonymous at July 21 2025, 10:59:19 AM.

SPARK JOB (http://10.0.2.15:4040/jobs/job?id=3) FINISHED

```
%spark
/*
Poblar la tabla con URL complexity:
*/

val dataWithComplexityUrl = data.withColumn(
    "UrlComplexity",
    getUrlComplexityUdf(
        col("NoOfSubDomain"),
        col("URLLength"),
        col("NoOfOtherSpecialCharsInURL"),
        col("SpacialCharRatioInURL"),
    )
)

dataWithComplexityUrl.select($"URL", $"UrlComplexity").where($"UrlComplexity" === "VERY COMPLEX").show(false)
```

```
+-------------------------------------------------------------------------------------------------------------------------------------------------------------
+-------------+
|URL                                                                                                                                                          |U
rlComplexity|
+-------------------------------------------------------------------------------------------------------------------------------------------------------------
+-------------+
|https://www.town.minamichita.lg.jp                                                                                                                           |M
EDIUM       |
|https://ipfs.io/ipfs/qmrvvyr84esa2assw9vvwupqjgsdn4c3dwkusfdwzdz3kn?clientid=noc@protocol.ai                                                                  |M
EDIUM       |
|http://att-103731-107123.weeblysite.com/                                                                                                                     |M
EDIUM       |
|https://fb-restriction-case-97be5.web.app/                                                                                                                   |M
EDIUM       |
|https://pontosapontamentolu.com/gclid/=/c/?gclid=ishecabh95cvbhb1eiwagy6m4vn7url3tlj5m_nu__lcyj4n06pqquydbh-56148buwwbmv7k1qejboci-isjw5_kiq|M
EDIUM       |
|https://s3.amazonaws.com/appforest_uf/f1678949673383x832048620362898600/index%20%284%29.html                                                                |M
```

Took 2 sec. Last updated by anonymous at July 21 2025, 10:59:23 AM. (outdated)

Proyecto programación week14

SPARK JOB (http://10.0.2.15:4040/jobs/job?id=4) FINISHED

```
%spark
/*
Poblar la tabla con presencia de contenido
*/
val dataWithPresenceRatio = data.withColumn(
    "PresenceRatio",
    getPresenceRatioUdf(
        col("NoOfImage"),
        col("NoOfCSS"),
        col("NoOfJS"),
        col("LineOfCode"),
    )
)


dataWithPresenceRatio.select($"URL", $"PresenceRatio").show(false)
```

```
+--------------------------------+-------------+
|URL                             |PresenceRatio|
+--------------------------------+-------------+
|https://www.southbankmosaics.com  |LOW          |
```

```
|https://www.uni-mainz.de          |LOW          |
|https://www.voicefmradio.co.uk    |VERY LOW     |
|https://www.sfnmjournal.com       |VERY LOW     |
|https://www.rewildingargentina.org|VERY LOW     |
|https://www.globalrepertoire.com  |VERY LOW     |
|https://www.saffronart.com        |VERY LOW     |
|https://www.nerdscandy.com        |LOW          |
|https://www.hyderabadonline.in    |VERY LOW     |
|https://www.aap.org               |VERY LOW     |
|https://www.religionenlibertad.com|VERY LOW     |
|http://www.teramill.com           |VERY LOW     |
|https://www.socialpolicy.org      |VERY LOW     |
|https://www.aeb61.com             |VERY LOW     |
```

Proyecto programación week14

Took 2 sec. Last updated by anonymous at July 21 2025, 10:59:28 AM.

Proyecto programación week14

```
%spark

data.select($"Domain") // Selecciona solo la columna 'Domain'
  .where(col("Domain").endsWith(".ec")) // Filtra donde la columna 'Domain' contenga la subcadena "ec"
  .show(200, false)
```

SPARK JOB FINISHED

```
+----------------------------------+
|Domain                            |
+----------------------------------+
|www.eltelegrafo.com.ec            |
|www.esmil.mil.ec                  |
|www.udla.edu.ec                   |
|www.elmercurio.com.ec             |
|www.loja.gob.ec                   |
|www.ueea.edu.ec                   |
|stanford.edu.ec                   |
|www.utpl.edu.ec                   |
|www.cancilleria.gob.ec            |
|www.registrocivil.gob.ec          |
|www.gestionderiesgos.gob.ec       |
|sotein.com.ec                     |
|www.job.ec                        |
|www.uazuay.edu.ec                 |
```

Took 1 sec. Last updated by anonymous at July 15 2025, 5:36:56 PM.

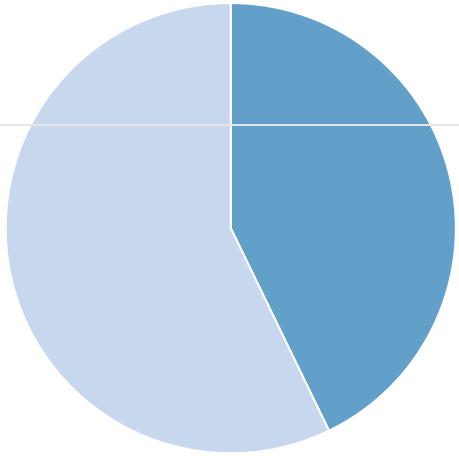
```
%spark
// Gráfica para saber cuantas páginas son de phishing y cuales no lo son
val phishingPages = data.select($"label")
.groupBy($"label".as("Legítima"))
.agg(count("*").as("Total"))


z.show(phishingPages)
```

SPARK JOB FINISHED

settings

Proyecto programación week1



```
phishingPages: org.apache.spark.sql.DataFrame = [Legítima: int, Total: bigint]
```

Took 3 sec. Last updated by anonymous at July 22 2025, 8:14:01 PM. (outdated)

Consultas en base a las preguntas

FINISHED

Las siguientes consultas tienen como objetivo responder a las preguntas planteadas previamente.

Took 4 sec. Last updated by anonymous at July 21 2025, 10:58:05 AM.

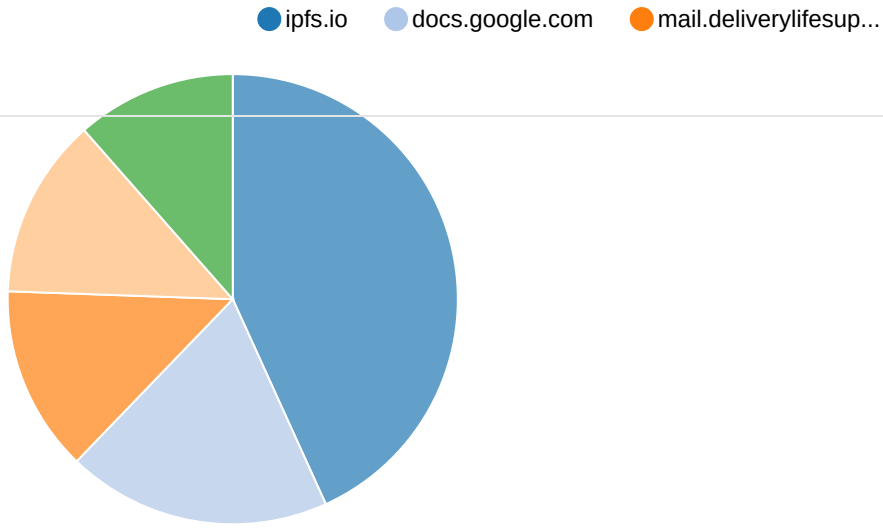
```
%spark
// ¿Cuáles son los cinco dominios más recurrentes en el dataset?
val fiveDomains = data.select($"domain").groupBy($"domain")
.agg(count($"domain").as("Total dominios"))
.orderBy($"Total dominios".desc)
.limit(5)

z.show(fiveDomains)
```

SPARK JOB (http://10.0.2.15:4040/jobs/job?id=11) FINISHED

settings

Proyecto programación week1




```
fiveDomains: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [domain: string, Total dominios: bigint]
```

Took 4 sec. Last updated by anonymous at July 21 2025, 11:10:31 AM.

SPARK JOB FINISHED

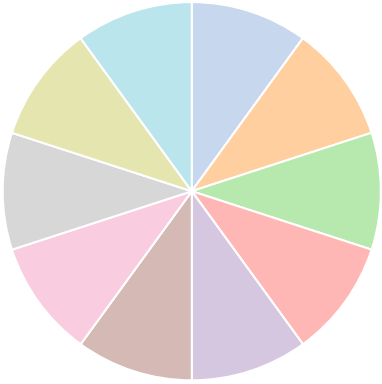
```
%spark
// ¿Cuáles son los factores más comunes en las páginas consideradas como phishing?
val dangerousPages = data.select($"URL", $"label", $"NoOfURLRedirect", $"Crypto", $"HasCopyrightInfo", $"IsHTTPS")
.where($"NoOfURLRedirect" > 0)
.where($"Crypto" === 1)
.where($"HasCopyrightInfo" === 0)
.where($"IsHTTPS" === 0)


z.show(dangerousPages)
```

settings

http://crypto-nodes.... http://crypto-nodes.... http://mycrypto-regu... http://mycrypto-regu... http://exodususer.co... http://exodususer.co... htt
http://cryptomgnode... http://cryptonodes.n... http://www.metamask1... http://www.metamask1... http://www.strategie... http://www.strategie... htt
http://www.cryptokoi... http://www.cryptokoi... http://atendimentobt... http://atendimentobt...

Proyecto programación week14



```
dangerousPages: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [URL: string, label: int ... 4 more fields]
```

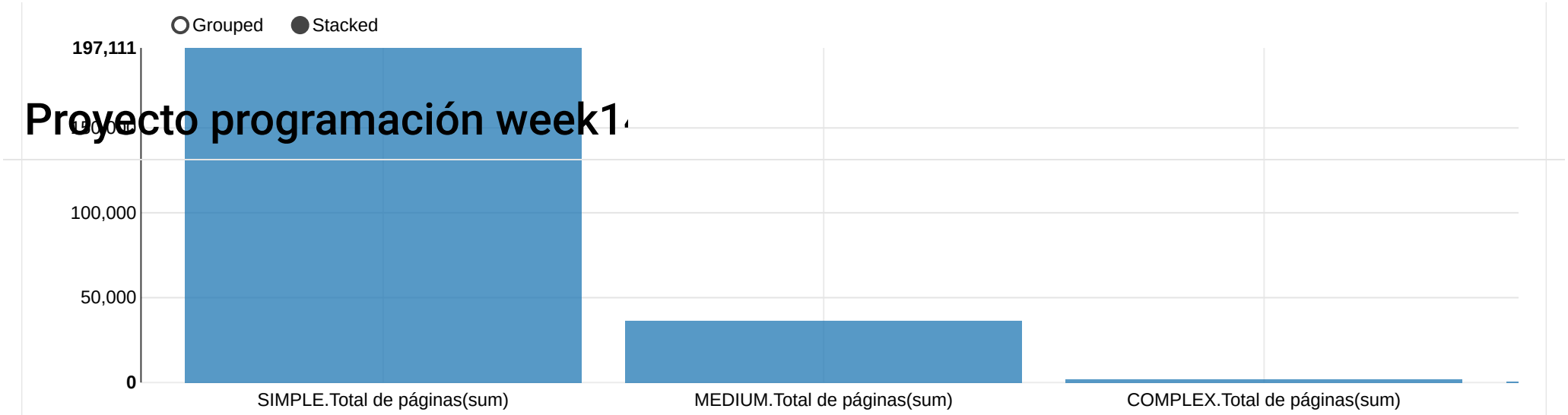
Took 3 sec. Last updated by anonymous at July 22 2025, 8:19:03 PM. (outdated)

```
%spark
//¿Cuántas páginas de cada tipo de complejidad de URL existen dentro del dataset?
val dataUrlComplexity = dataWithComplexityUrl.select($"URL", $"UrlComplexity")
.groupBy($"UrlComplexity")
.agg(count($"UrlComplexity").as("Total de páginas"))
.orderBy($"Total de páginas".desc)

z.show(dataUrlComplexity)
```

SPARK JOB (http://10.0.2.15:4040/jobs/job?id=32) FINISHED

settings




```
dataUrlComplexity: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [UrlComplexity: string, Total de páginas: bigint]
```

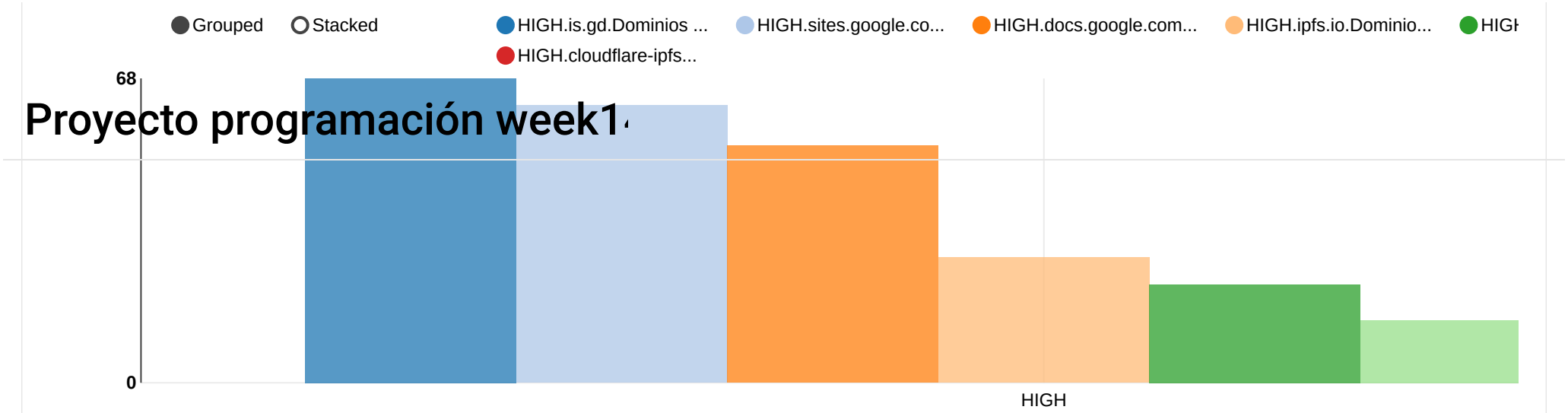
Took 2 sec. Last updated by anonymous at July 21 2025, 11:28:26 AM. (outdated)

```
%spark
//¿Cuáles son los cinco dominios con más seguridad basados en la columna derivada?
val highSecurityDomains = dataWithSecurity.select($"SecurityCategory", $"Domain")
.where($"SecurityCategory" === "HIGH")
.groupBy($"SecurityCategory", $"Domain")
.agg(count("*").as("Dominios totales"))
.where($"Dominios totales" > 10)
.orderBy($"Dominios totales".desc)

z.show(highSecurityDomains)
```

SPARK JOB (http://10.0.2.15:4040/jobs/job?id=35) FINISHED

settings




```
highSecurityDomains: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [SecurityCategory: string, Domain: string ... 1 more field]
```

Took 2 sec. Last updated by anonymous at July 21 2025, 11:39:36 AM. (outdated)

SPARK JOB FINISHED

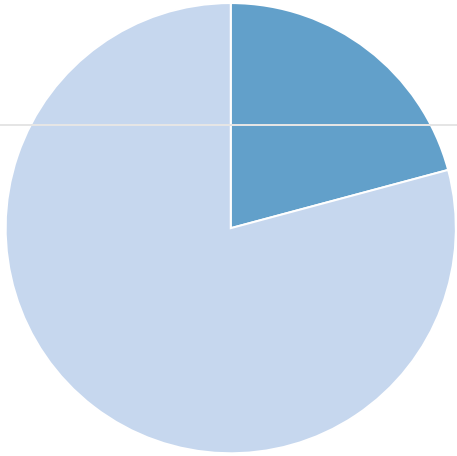
```
%spark
//¿Cuántas páginas con dominio ‘ec’ existen dentro del dataset y que porcentaje de estas con consideradas como ‘phishing’?

val ecuadorDomains = data.select($"label".as("legitima"))
  .where(col("Domain").endsWith(".ec"))
  .groupBy($"legitima")
  .agg(count("*").as("Total dominios .ec"))

  z.show(ecuadorDomains)
```

settings

Proyecto programación week1



```
ecuadorDomains: org.apache.spark.sql.DataFrame = [legitima: int, Total dominios .ec: bigint]
```

Took 2 sec. Last updated by anonymous at July 21 2025, 11:46:15 AM. (outdated)

```
%spark
data.count()
```

SPARK JOB (http://10.0.2.15:4040/jobs/job?id=6) FINISHED

```
res4: Long = 235795
```

Took 2 sec. Last updated by anonymous at July 10 2025, 7:12:16 PM.

```
%spark
```

READY