**Vinicio Mora / Daniela Valerio / Benjamín Calvo**

Prof. Lorena Zúñiga

**Instituto Tecnológico de Costa Rica.  
Administración de Tecnologías de Información.**

Proyecto #2

TI-8507 Evaluación de la Calidad de Datos



Contenido

[Introducción 3](#_Toc453711306)

[Perfilado de Atributos 5](#_Toc453711307)

[**Tabla Compras** 5](#_Toc453711308)

[Estadísticas resumen 11](#_Toc453711309)

[**Tabla Tiendas** 14](#_Toc453711310)

[**Tabla Clientes** 24](#_Toc453711311)

[**Tabla Productos** 32](#_Toc453711312)

[Agregación de Patrones 40](#_Toc453711313)

[**IDCLIENTE** 40](#_Toc453711314)

[**CODTIENDA** 41](#_Toc453711315)

[**CODPROD** 42](#_Toc453711316)

[**FECNAC** 44](#_Toc453711317)

[**PRODFEC** 45](#_Toc453711318)

[Reglas de Negocio 47](#_Toc453711319)

[**Regla fecha de compra** 47](#_Toc453711320)

[**Regla Mayoría de Edad** 50](#_Toc453711321)

[Perfilado de Relaciones. 54](#_Toc453711322)

[**COMPRAS-CLIENTES** 54](#_Toc453711323)

[**COMPRAS-TIENDAS** 56](#_Toc453711324)

[**TIENDAS-CANTON** 57](#_Toc453711325)

[**TIENDAS-DISTRITO** 59](#_Toc453711326)

[Restricciones de valores válidos 61](#_Toc453711327)

[Tabla: Cantones 61](#_Toc453711328)

[Screenshot 62](#_Toc453711329)

[Tabla: Clientes 62](#_Toc453711330)

[Screenshot 63](#_Toc453711331)

[Tabla: Compras 63](#_Toc453711332)

[Screenshot 64](#_Toc453711333)

[Tabla: Distritos 64](#_Toc453711334)

[Screenshot 65](#_Toc453711335)

[Tabla: Productos 65](#_Toc453711336)

[Screenshot 66](#_Toc453711337)

[Tabla: TIENDAS 66](#_Toc453711338)

[Screenshot 67](#_Toc453711339)

[Restricciones de formato 67](#_Toc453711340)

[Tabla: Cantones 67](#_Toc453711341)

[Screenshot 68](#_Toc453711342)

[Tabla: Clientes 68](#_Toc453711343)

[Screenshot 69](#_Toc453711344)

[Tabla: Compras 69](#_Toc453711345)

[Screenshot 70](#_Toc453711346)

[Tabla: Distritos 70](#_Toc453711347)

[Screenshot 71](#_Toc453711348)

[Tabla: Productos 71](#_Toc453711349)

[Screenshot 72](#_Toc453711350)

[Tabla: Tiendas 72](#_Toc453711351)

[Screenshot 73](#_Toc453711352)

[Restricciones de referencia 73](#_Toc453711353)

[Tabla: Tiendas y provincia 73](#_Toc453711354)

[Screenshot 74](#_Toc453711355)

[Tabla: Compras y cliente 74](#_Toc453711356)

[Screenshot 74](#_Toc453711357)

[Dependencia parcial 75](#_Toc453711358)

[Tabla: Provincia-Cantones 75](#_Toc453711359)

[Screenshot 76](#_Toc453711360)

[Tabla: Provincia-Cantones-Distritos 76](#_Toc453711361)

[Screenshot 77](#_Toc453711362)

[Restricción de identidad 77](#_Toc453711363)

[Screenshot 78](#_Toc453711364)

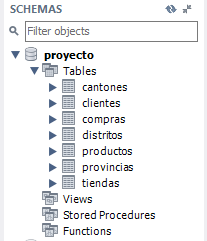
[Descripción del Diseño de la base de datos del Catalogo 79](#_Toc453711365)

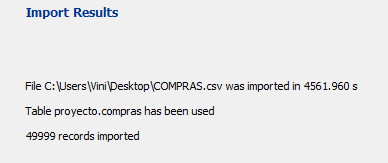
[Diseño y funcionalidades de la aplicación del Catálogo de Reglas. 79](#_Toc453711366)

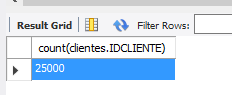
# **Introducción**

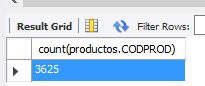
El presente proyecto consiste en el diseño e implementación de reglas de calidad de datos a un conjunto de tablas suministradas, además de la creación de una aplicación que sirva como catálogo de las reglas diseñadas para la evaluación de la calidad en los datos.

Para la realización de este proyecto se importaron los datos suministrados en una base de datos MySQL, esto debido a la facilidad con que permite para la importación de Datasets y para su conexión con Talend.

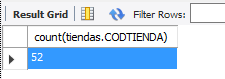
Para la importación de los datos se creó la base de datos “proyecto”, los datos suministrados fueron importados de manera correcta y en su totalidad.

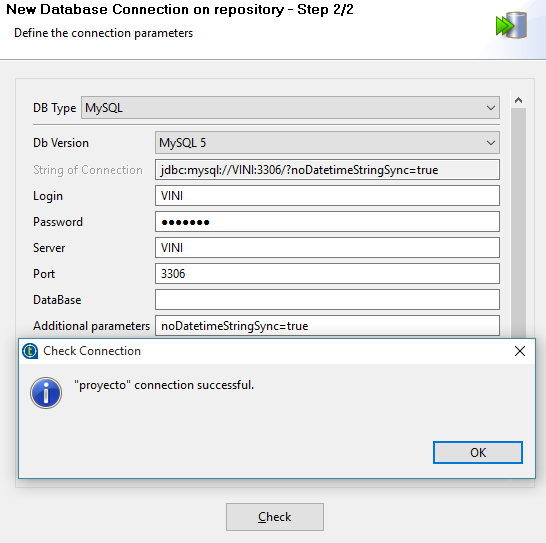
La tabla compras se importó con 49999 registros.

La tabla de clientes se importó con los 25000 registros.

La tabla productos se importó con los 3625 registros.

La tabla tiendas se importó con los 51 registros.



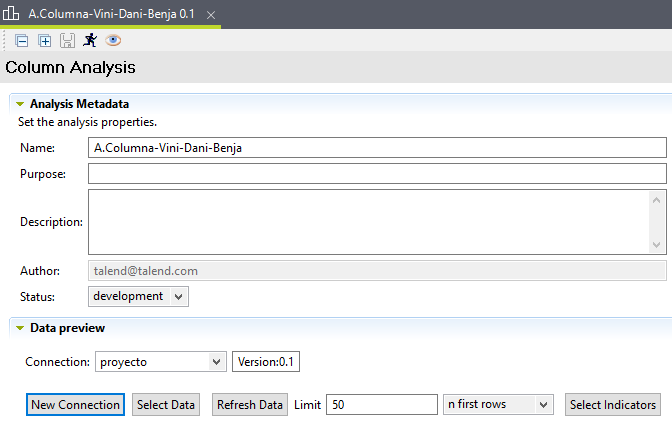
La conexión de la base de datos “proyecto” con Talend se efectuó de manera exitosa

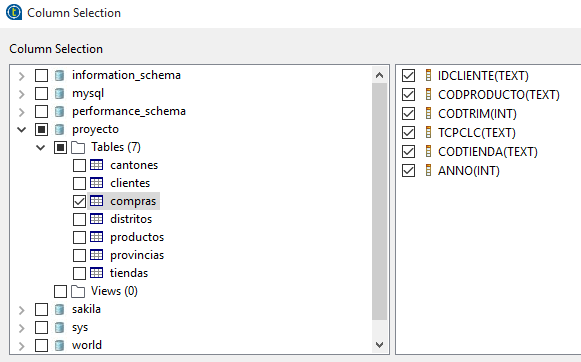
# **Perfilado de Atributos**

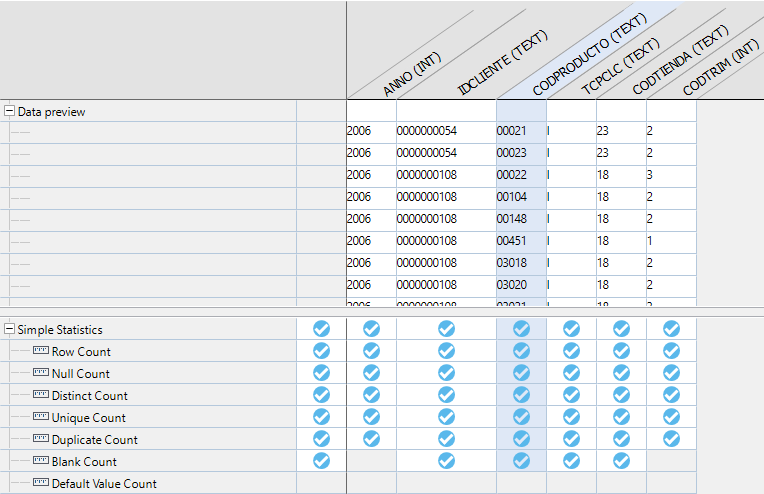
En la siguiente sección se muestra los resultados del perfilado de atributos sobre cada uno de los atributos de las tablas Compras, Tiendas, Clientes y Productos

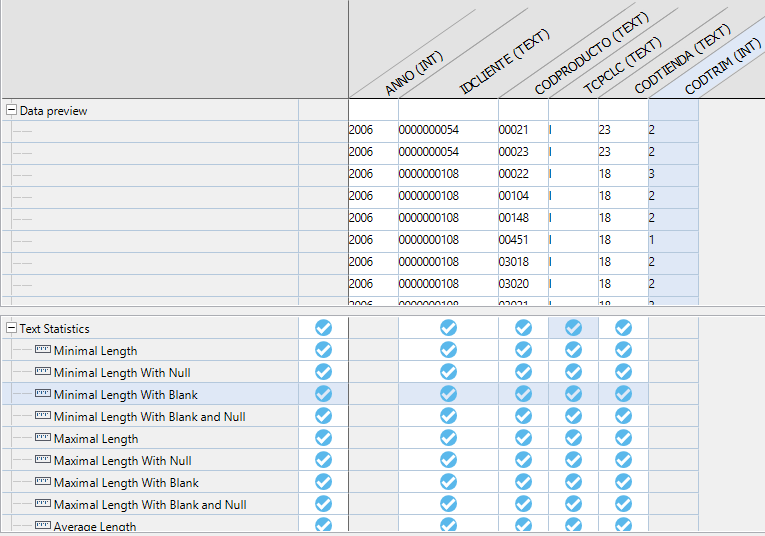
## **Tabla Compras**

**Análisis Básico de columna**

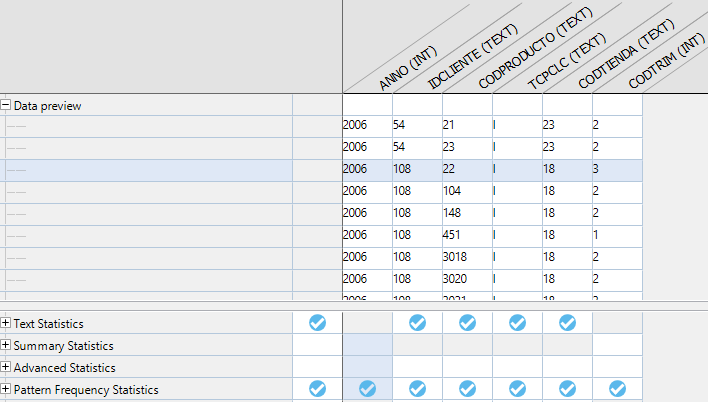
Se crea el análisis bajo el nombre A.Columna-Vini-Dani-Benja.

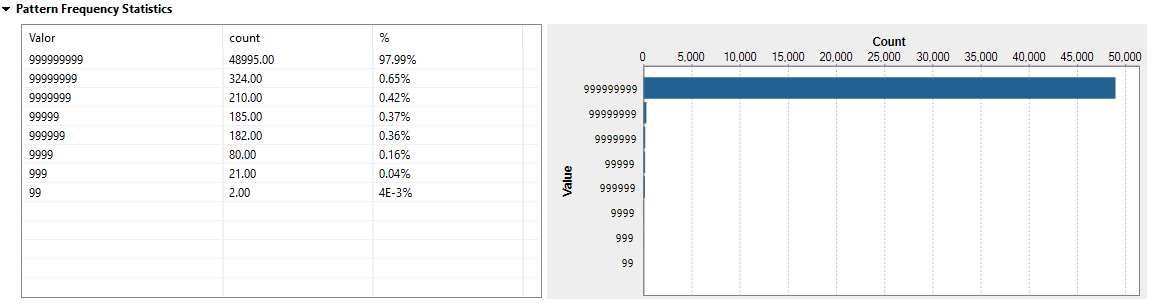
Se seleccionan los 6 indicadores de la tabla Compras.

Se seleccionan los indicadores y se elige “Estadísticas Simples” para todos los atributos.

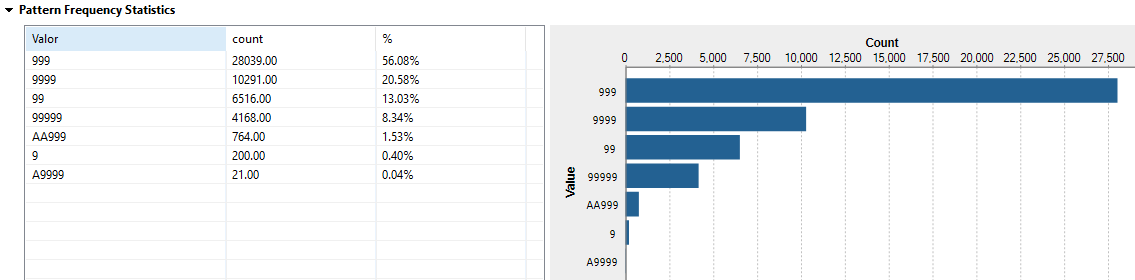
Se selecciona “Estadísticas de Texto” para este tipo de atributos.

**Análisis de Frecuencia de Patrones**

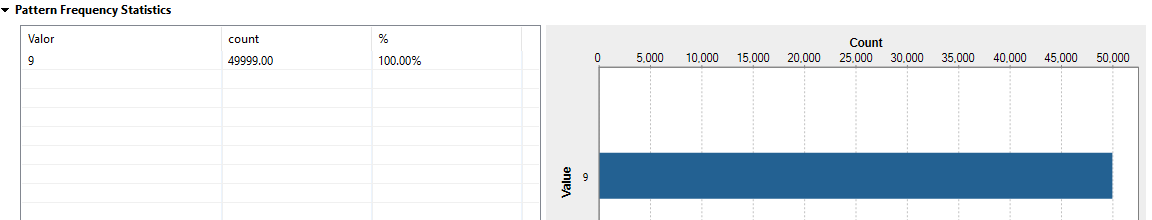
Se realiza el análisis de frecuencia de patrones.

A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo IDCLIENTE.

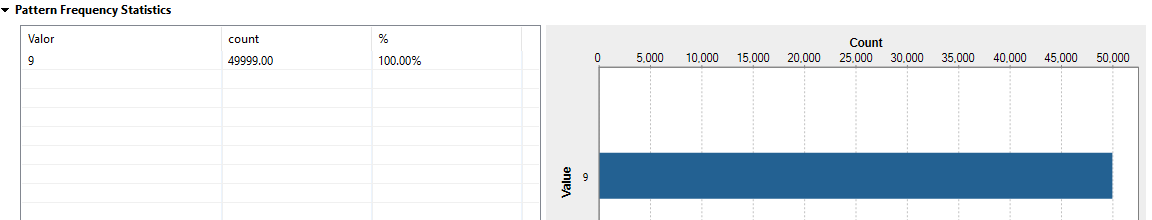
En este se puede observar como el 97.99% de los registros tienen 9 dígitos en este atributo. Todos los atributos tienen formato correcto ya que van de los 2 a los 9 dígitos.

A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo CODPRODUCTO.

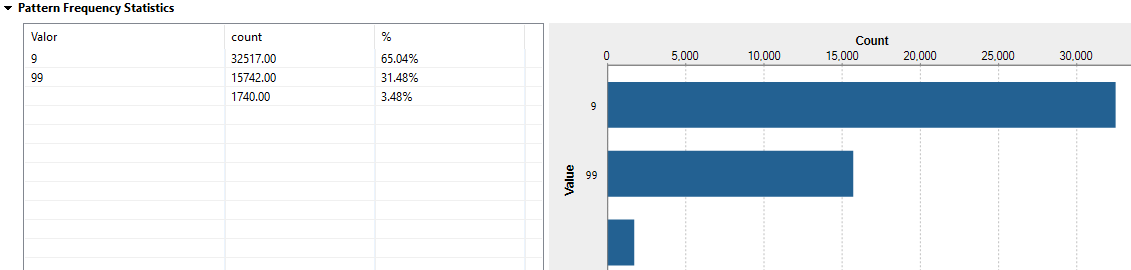
En este se nota como el 56% de los registros tienen 3 caracteres numéricos en este atributo y el 0.40% solo tienen un carácter.

A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo CODTRIM.

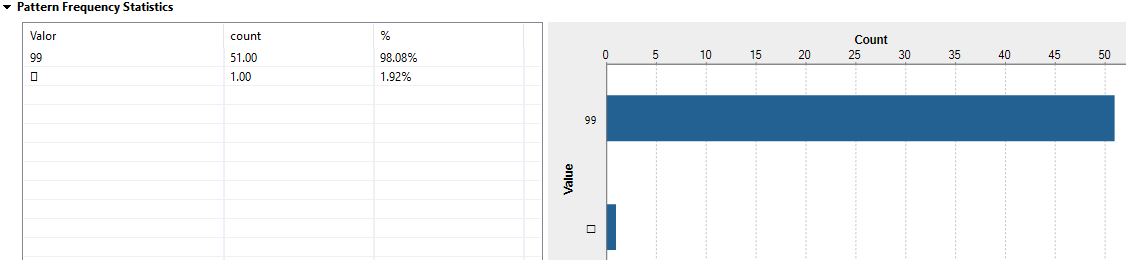
Se nota como el 100% de los atributos tienen un único carácter. Lo cual es el formato correcto.

A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo TCPCLC.

Al igual que en el caso anterior el 100% cuenta con un único carácter numérico. Lo que corresponde a un formato válido.

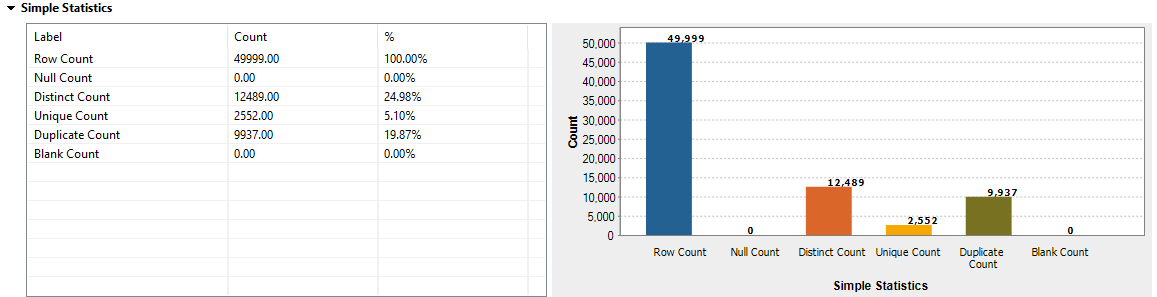
A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo CODTIENDA.

Este análisis muestra que existe un 3.48% de registros que no poseen este atributo. Existe un 65.04% de valores con un único valor numérico lo cual es inválido ya que se estipula que deben ser dos dígitos.

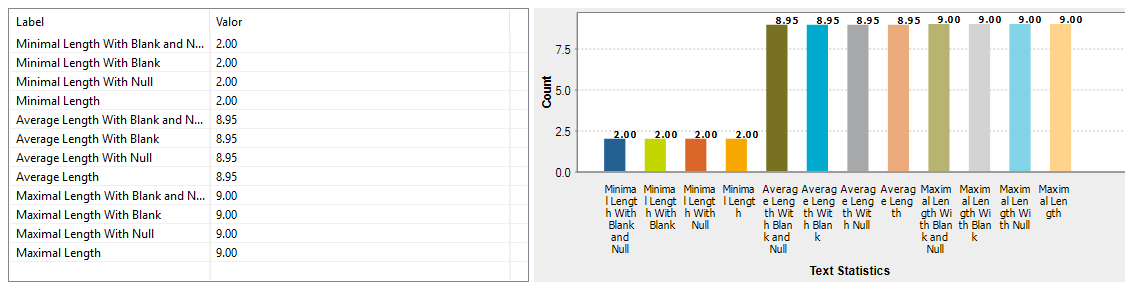
A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo CODTIENDA.

Esta muestra que el 99% corresponde al valor válido de 2 dígitos.

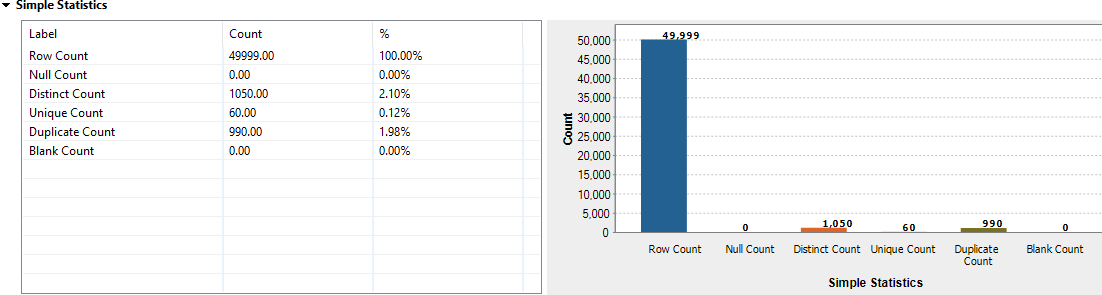
# **Estadísticas resumen**

**Estadísticas Simples para el atributo IDCLIENTE**

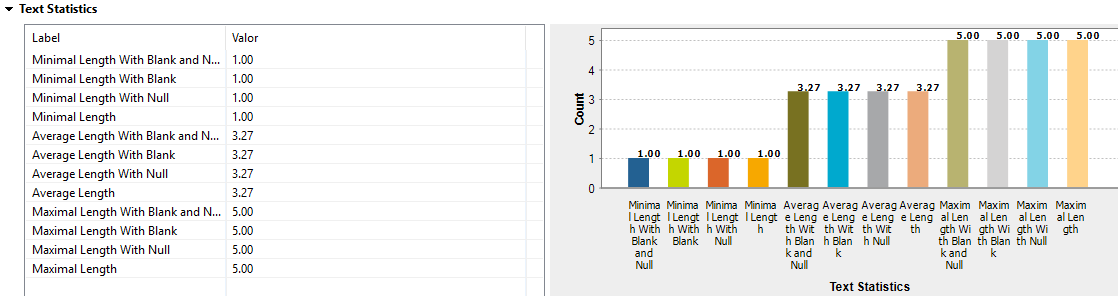
Este análisis muestra un 19.67% de valores duplicados, lo cual indica varias compras realizadas por un cliente.

**Estadísticas textuales para el atributo IDCLIENTE**

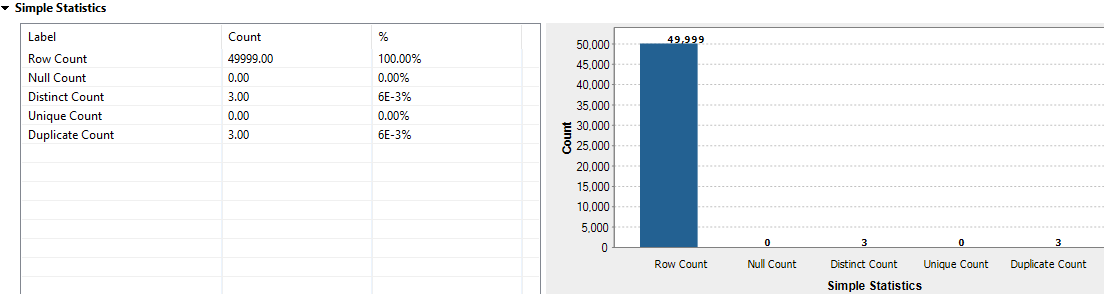
En este los valores mínimos son de 2.00, los medios de 8.95 y los máximos de 9.00.

**Estadísticas Simples para el atributo CODPRODUCTO**

En este análisis se muestra un 1.98% de registros duplicados, lo que indica compras de un mismo producto.

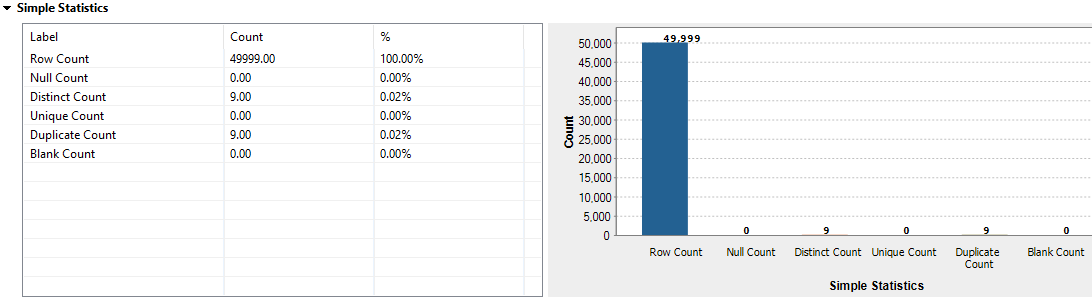
**Estadísticas textuales para el atributo CODPRODUCTO**

Este análisis muestra valores mínimos en blanco y nulos de 1.00, medios de 3.27 y máximos de 5.00.

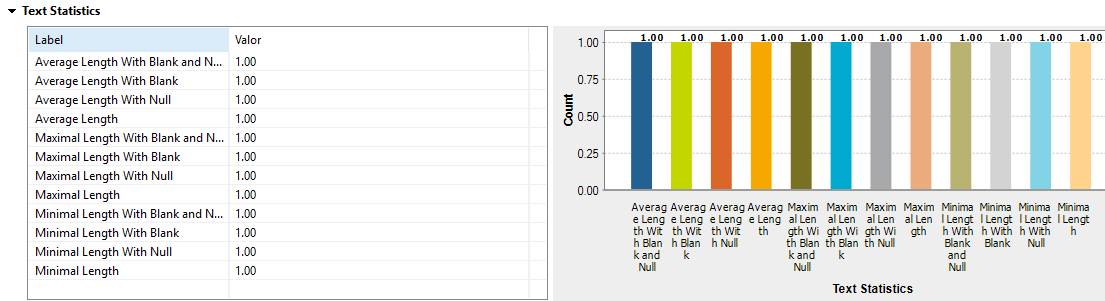
**Estadísticas Simples para el atributo CODTRIM**

Este análisis muestra un total de 3 registros duplicados, lo cual corresponde a los 3 trimestres 1,2 y 3, como se muestra a continuación

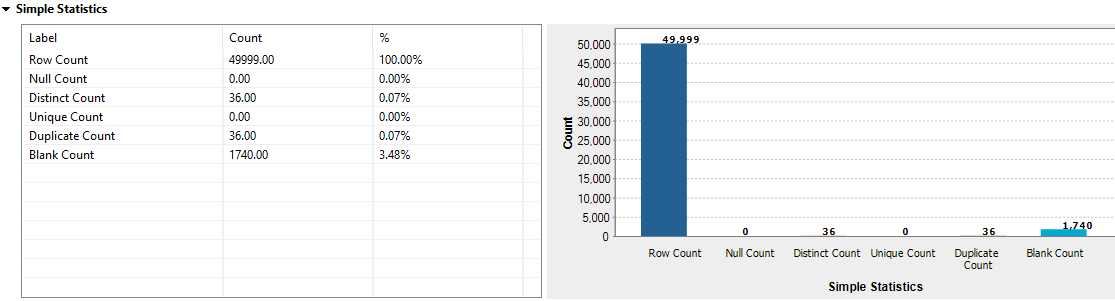
.

**Estadísticas Simples para el atributo TCPCLC**

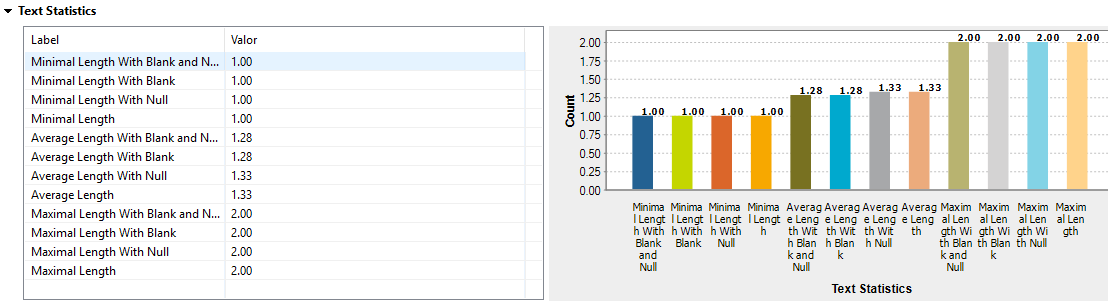
Este análisis muestra un 0.02% de registros duplicados, que serían los 9 valores para tipos de clientes.

**Estadísticas textuales para el atributo TCPCLC**

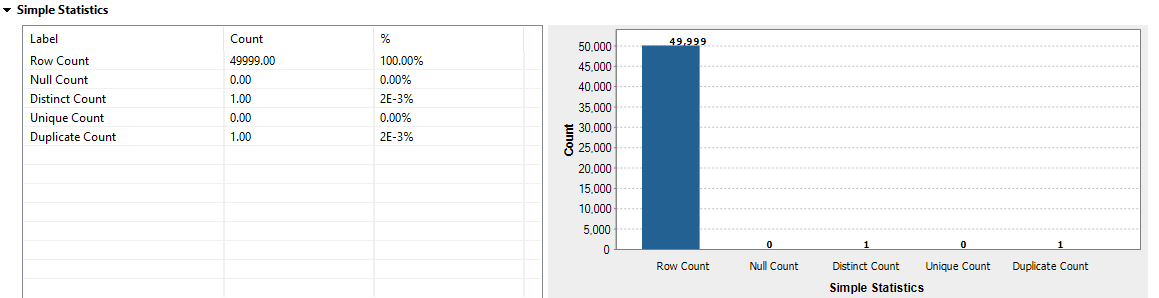
En este análisis se muestran valores de 1.00.

**Estadísticas Simples para el atributo CODTIENDA**

En esta se muestra un conteo de 3.48% de valores en blanco y 0.07% duplicados.

**Estadísticas textuales para el atributo CODTIENDA**

Los valores mínimos en blanco y nulos son de 1.00, los medios de 1.28, 1.20 y 1.33, los valores máximos son de 2.00.

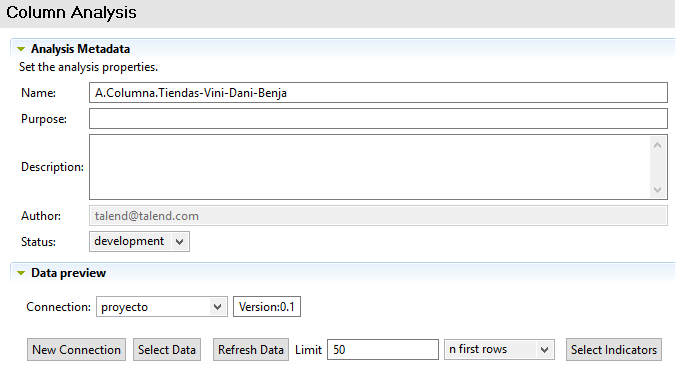
**Estadísticas Simples para el atributo ANNO**

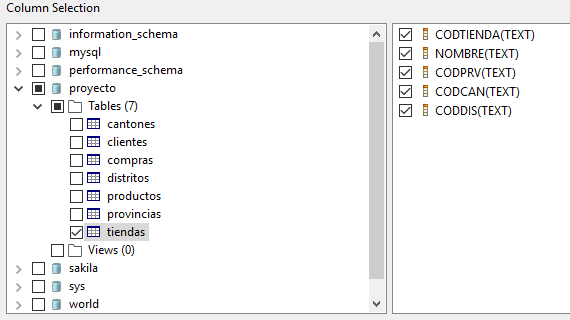
Se muestra solo 1 valor duplicado el cual corresponde al año 2006, ya que los 50.000 registros tienen este año como año como compra.

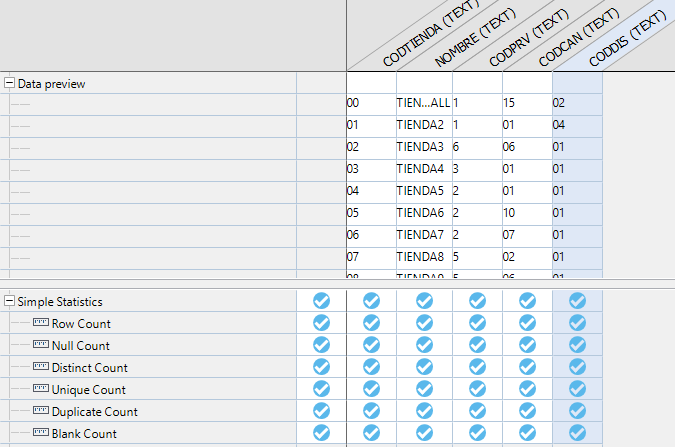
## **Tabla Tiendas**

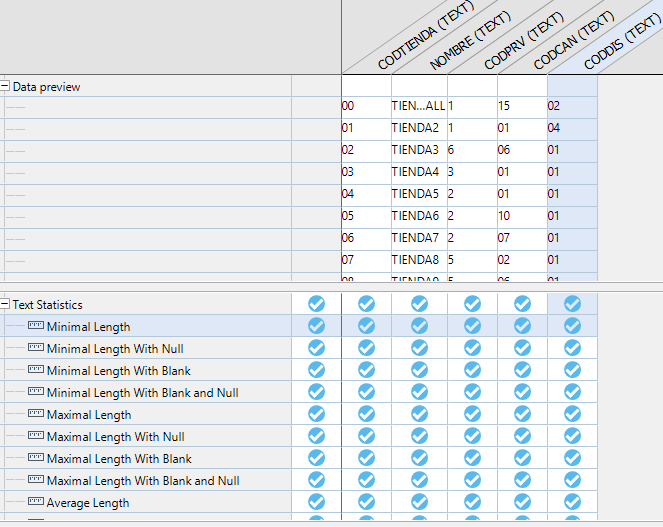
**Análisis Básico de columna**

Se crea el análisis bajo el nombre A.Columnas.Tiendas-Vini-Dani-Benja

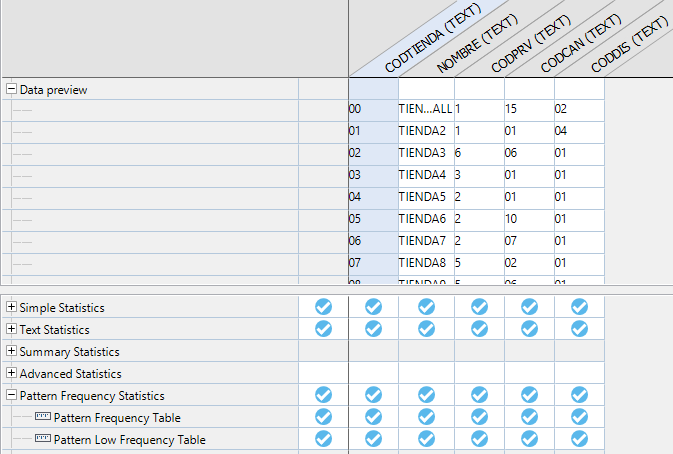


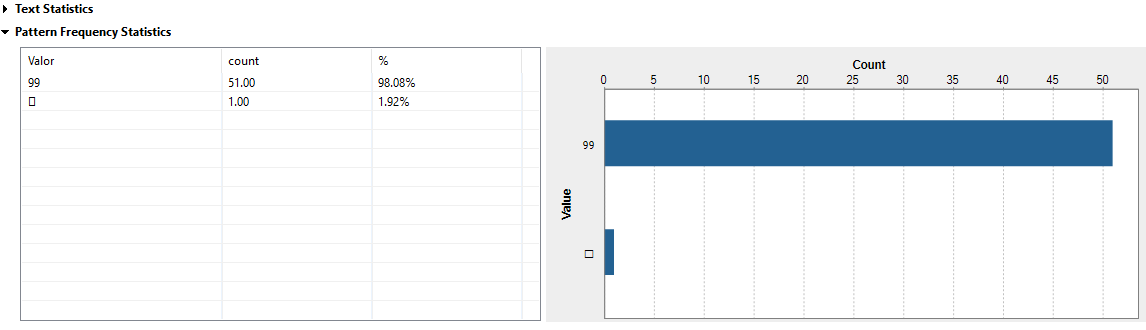
Se selecciona los indicadores de la tabla

Se seleccionan los indicadores y se elige “Estadísticas Simples” para todos los atributos.

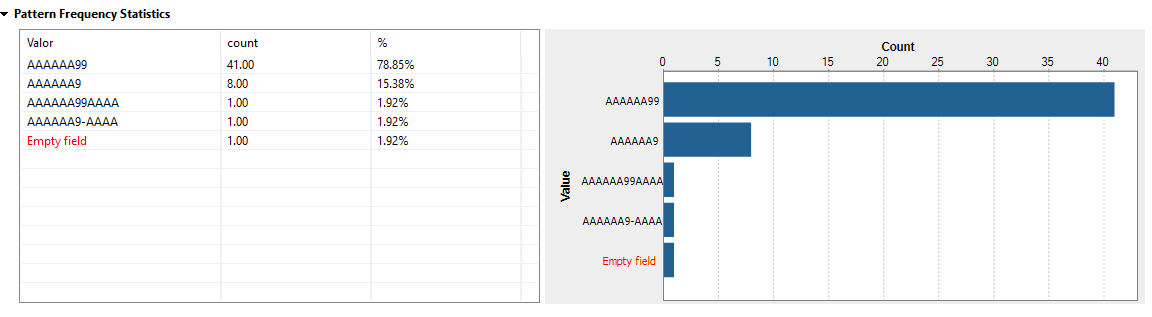
Se selecciona “Estadísticas de Texto” para este tipo de atributos.

**Análisis de Frecuencia de Patrones**

Se realiza el análisis de frecuencia de patrones.

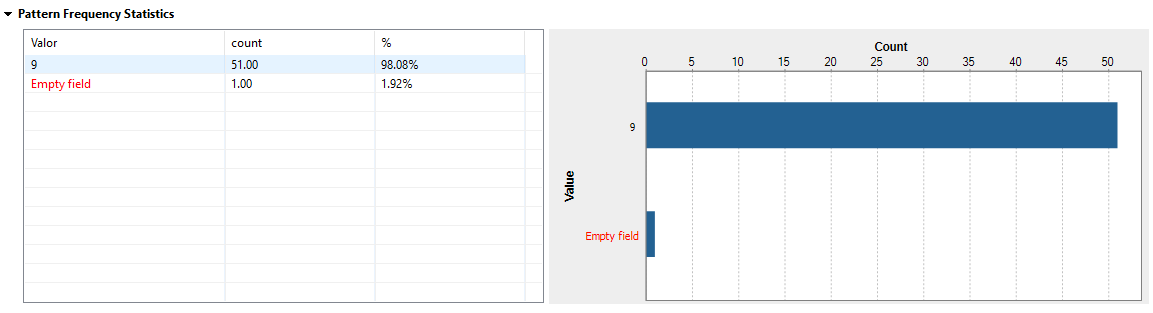
A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo CODTIENDA.

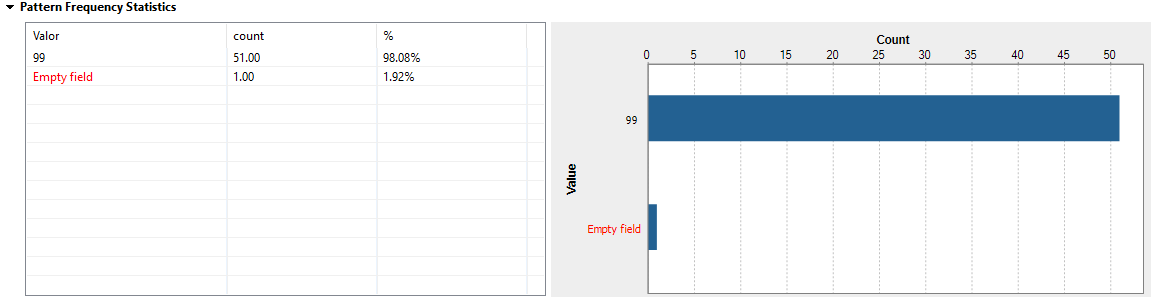
En este se muestra una frecuencia del 99.08% de valores de formato válidos (numéricos de 2 dígitos) contra un 1.92% de valores inválidos.

A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo Nombre.

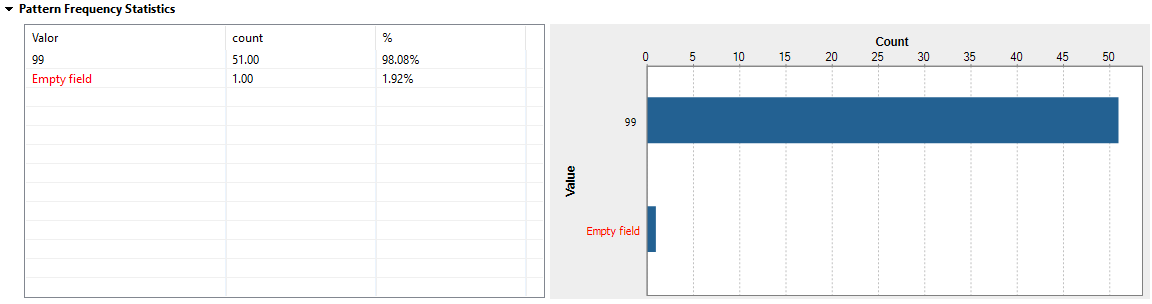
Este análisis al igual que el anterior muestra un 1.92% de espacios vacíos, los demás cumplen con el formato correcto.

A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo CODPRV.

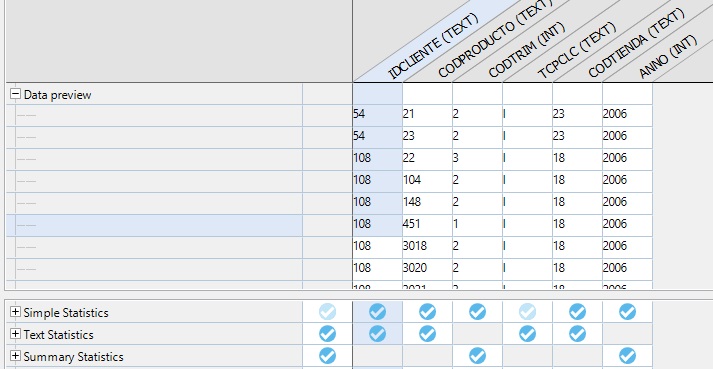
Este análisis muestra un 98.08% de valores correctos de 1 digito y un 1.92% vacíos.

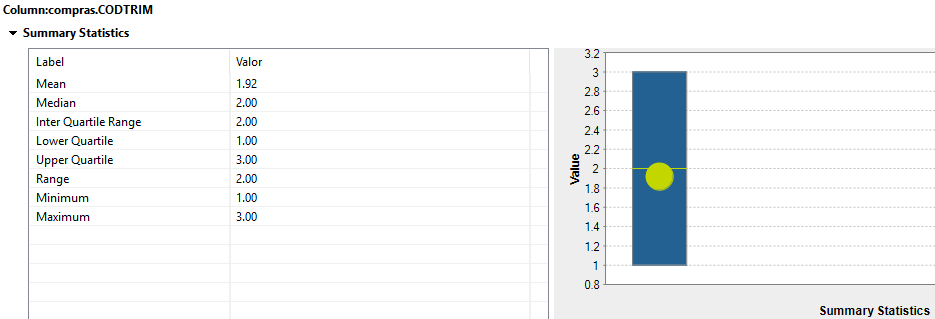
A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo CODCAN.

Este análisis muestra 98.08% valores de 2 dígitos, lo cual cumple el formato y un 1.92% de valores vacíos.

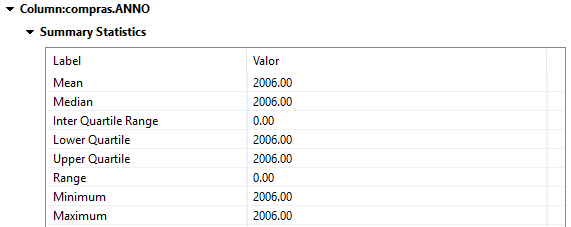
A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo CODDIS.

El análisis de esta columna también tiene como resultado un 98.08% de valores válidos y 1.92% vacíos.

El indicador “Estadísticas Resumen” necesita atributos de tipo int, por lo cual, solo aplica para la tabla Compras y sus atributos CODTRIM y ANNO.

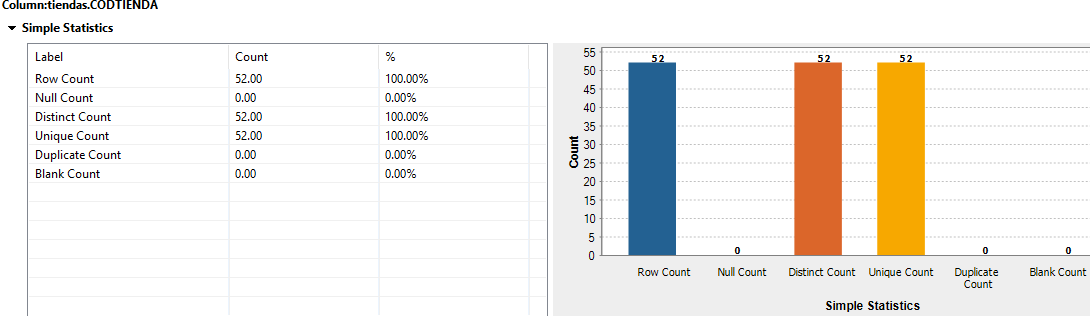
El resultado para el atributo CODTRIM es el siguiente:

En este se ve como los valores mínimos son de 1, que corresponde a los meses de enero, febrero y marzo y los máximos de 3 que corresponde a julio, agosto y setiembre.

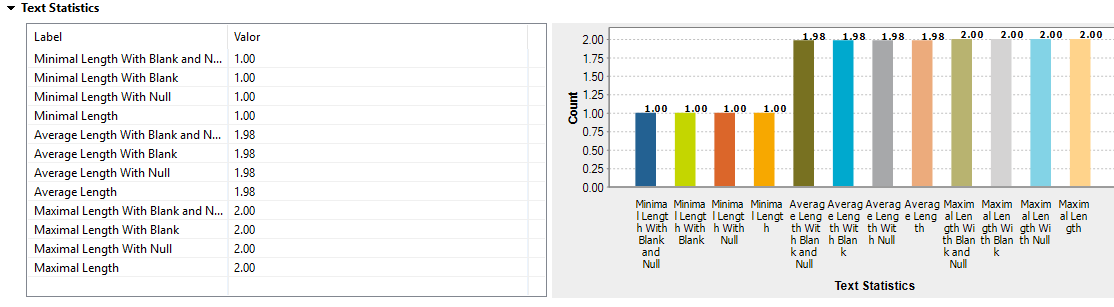
Las estadísticas resumen para el atributo año es la siguiente:

Se muestra el año 2006 en todos los rangos ya que es el año que predomina en las primeras 50.000 compras.

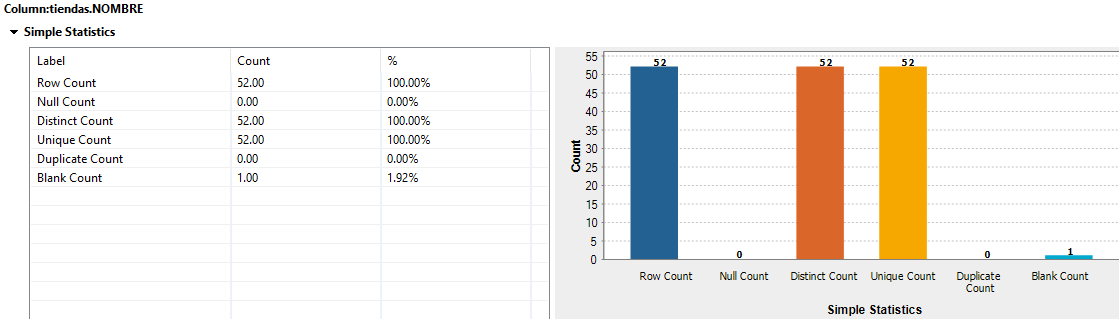
**Estadísticas resumen**

**Estadísticas Simples para el atributo CODTIENDA**

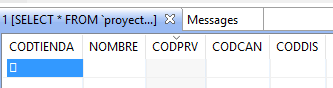
Este análisis muestra el conteo de la totalidad de las columnas, donde se muestra que no existen campos nulos, duplicados ni en blanco.

**Estadísticas textuales para el atributo CODTIENDA**

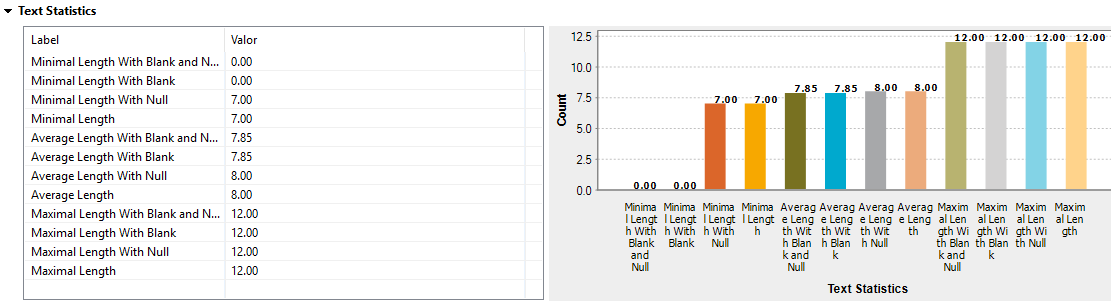
En este análisis se muestra como los valores mínimos tienen una media de 1.00, los valores promedio una media de 1.98 y los máximos una media de 2.00.

**Estadísticas simples para el atributo NOMBRE**

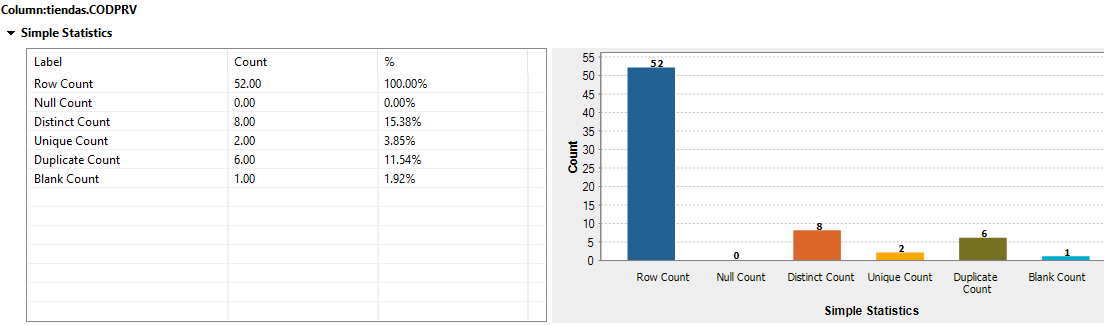
E en este análisis se muestra que, de la totalidad de los campos contados, no se encuentran valores nulos ni duplicados, pero si se encuentra un valor en blanco.

A continuación, se muestra el resultado de dicho campo en blanco:

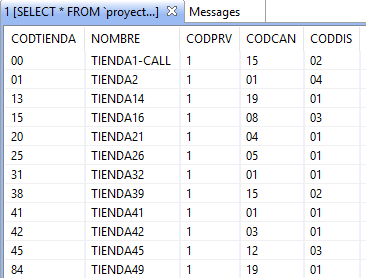
**Estadísticas textuales para el atributo NOMBRE**

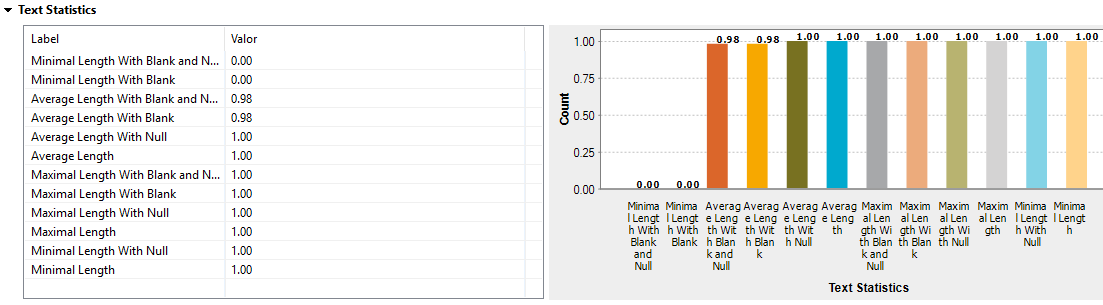


Se muestra como los valores mínimos con NULL son de 7.00, los valores promedio en blanco son de 7.85, los valores promedio nulos son de 8.00 y los valores máximos son de 12.00.

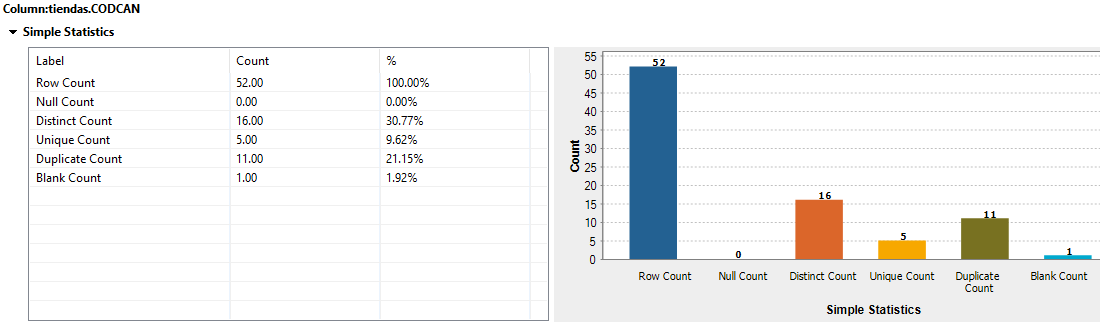
**Estadísticas Simples para el atributo CODPROV.**

En esta columna existe un porcentaje de 15.38% de campos distintos, un 3.85% de campos únicos, un 11.54% de campos duplicados y un 1.92% de campos en blanco.

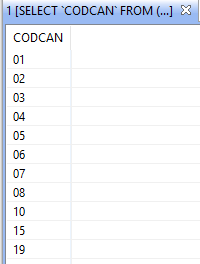
A continuación, se muestra el resultado de los campos duplicados correspondientes al 11.54%.

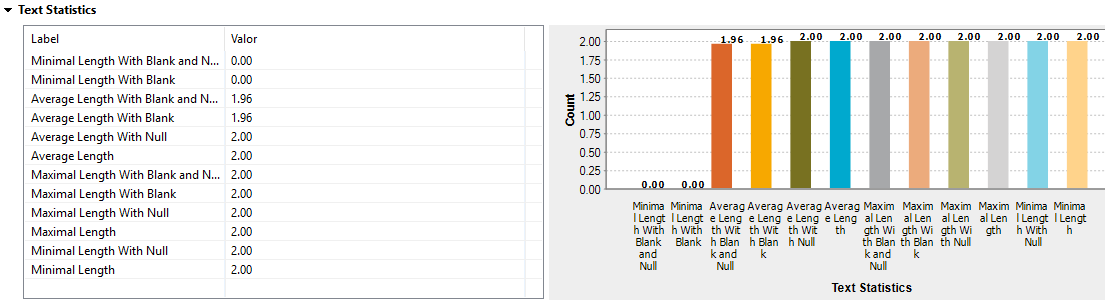
**Estadísticas textuales para el atributo CODPROV**

En este análisis se muestra como los valores mínimos son de 0.00, los medios para campos nulos y en blanco de 0.98, la media de nulos es un 1.00 y los valores máximos son de 1.00.

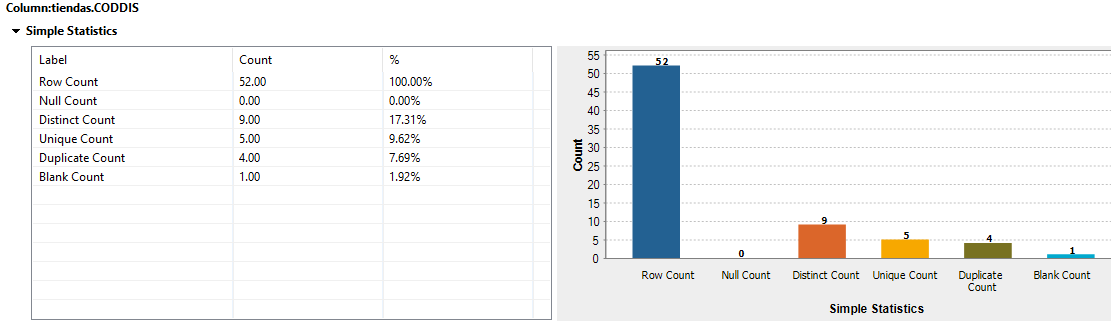
**Estadísticas Simples para el atributo CODCAN**

En este análisis se muestra la existencia de un 30.77% de campos distintos, 9.62% de campos únicos, 21.15% de campos duplicados y 1.92% de campos en blanco.

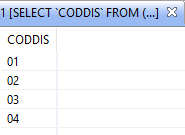
A continuación, se muestran los valores de los campos duplicados.

**Estadísticas textuales para el atributo CODCAN**

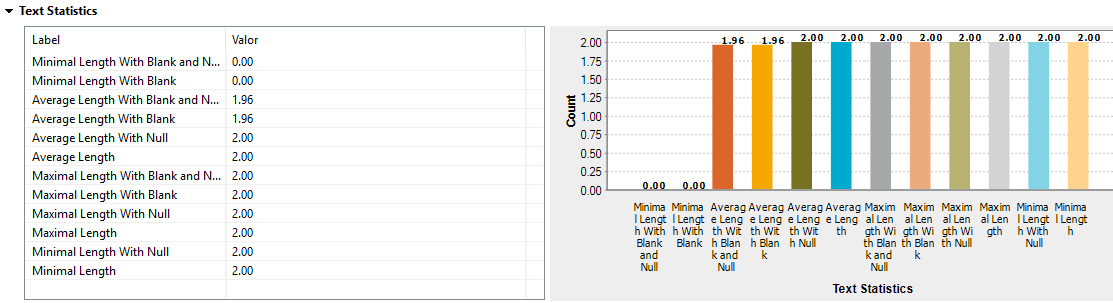
Se muestra como los valores en blanco mínimos son de 0.00, los promedio de 1.96, el promedio de nulos es de 2.00 y los máximos de 2.00.

**Estadísticas Simples para el atributo CODIS**

El resultado muestra un conteo de 9 valores distintos, 5 únicos, 4 duplicados y 1 en blanco.

A continuación, se muestran los valores duplicados.

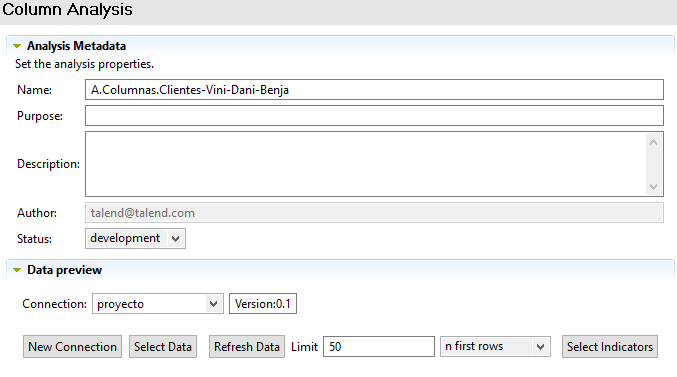
**Estadísticas textuales para el atributo CODIS**

Se muestra como los valores en blanco mínimos son de 1.00, los valores promedio son de 1.96. El promedio de nulos es de 2.00, y los máximos es de 2.00.

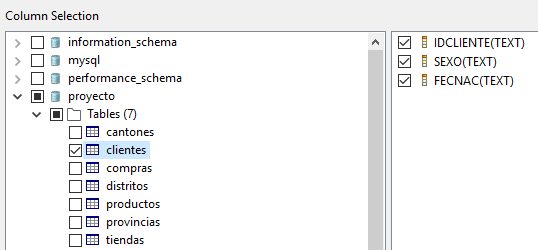
## **Tabla Clientes**

**Análisis Básico de columna**

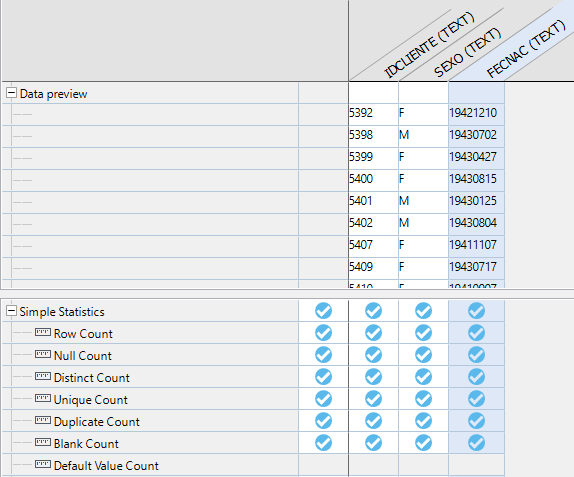
Se crea el análisis bajo el nombre A.Columnas.Clientes-Vini-Dani-Benja



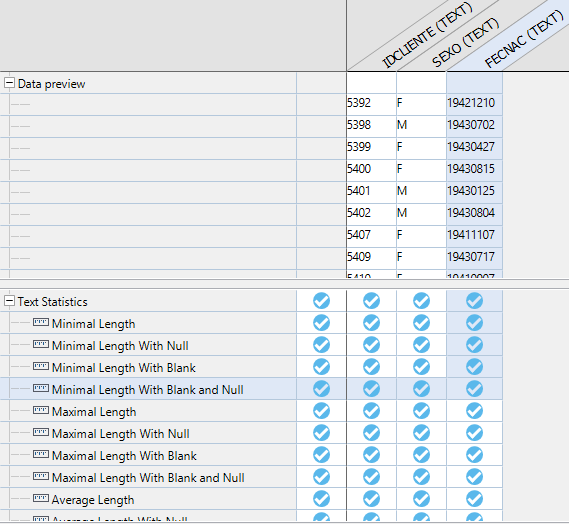
Se selecciona los indicadores de la tabla



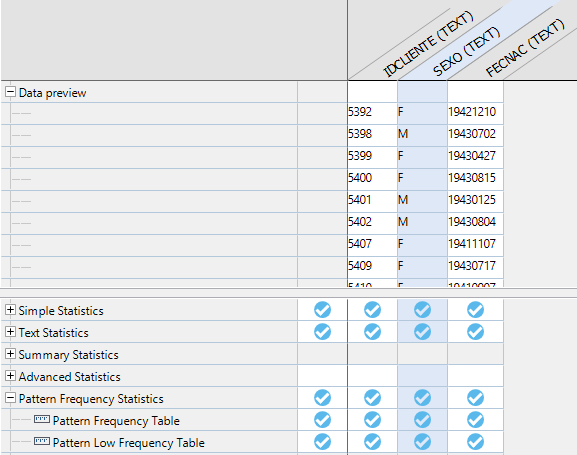
Se seleccionan los indicadores y se elige “Estadísticas Simples” para todos los atributos.

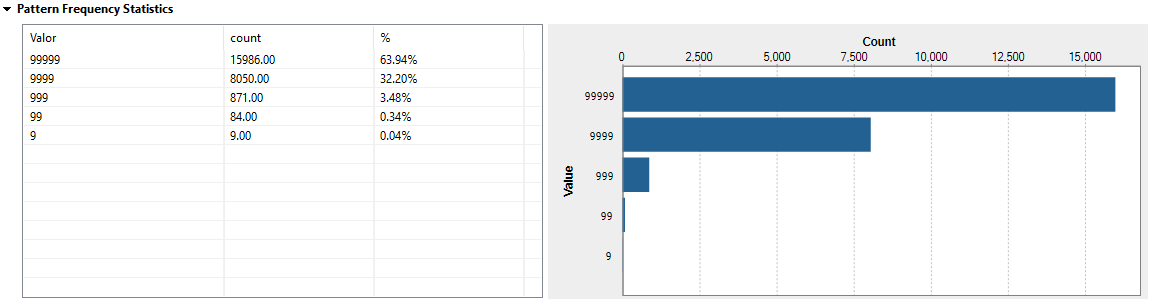


Se selecciona “Estadísticas de Texto” para este tipo de atributos.

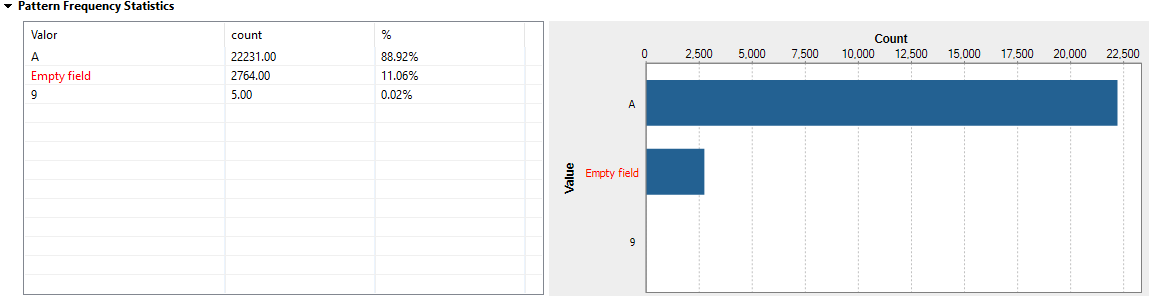


**Análisis de Frecuencia de Patrones**

Se selecciona el análisis de frecuencia para los campos IDCLIENTE, SEXO y FECNAC.

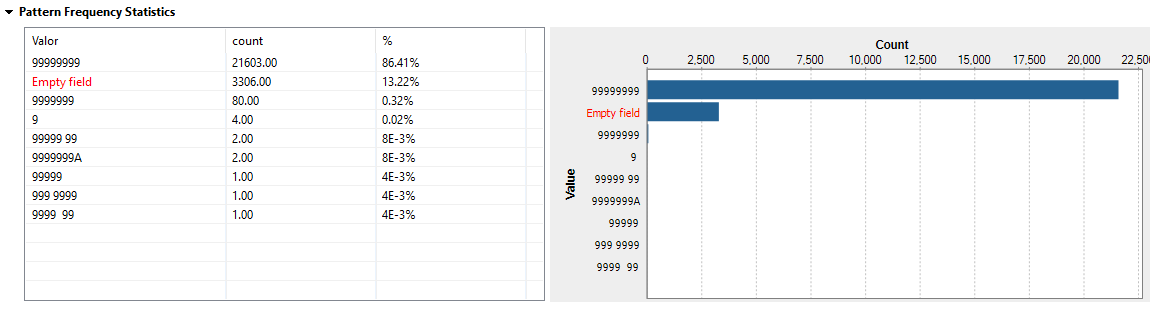
A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo IDCLIENTE.

En este se muestra que los códigos de cliente van de 1 a 5 dígitos, lo cual cumple el formato.

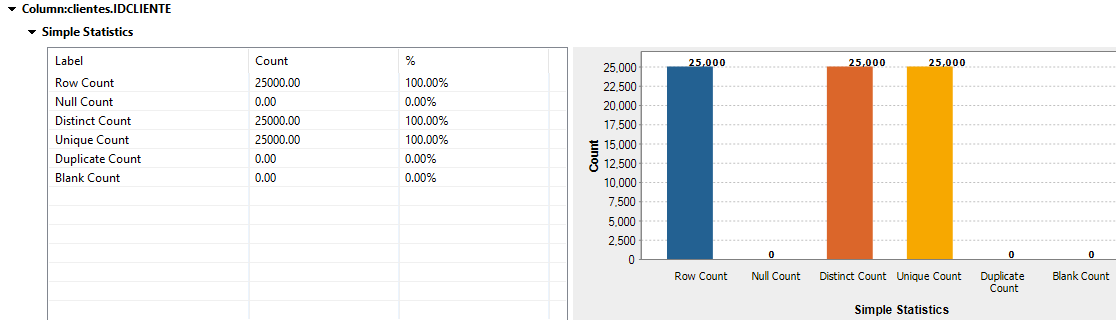
A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo SEXO.

En este análisis se refleja un total de 2764 campos vacíos lo que representa un 11.06% del total, además un 0.02% de valores inválidos en formato numérico, el resto posee un valor de carácter el cual pertenece a “M” o “F”.

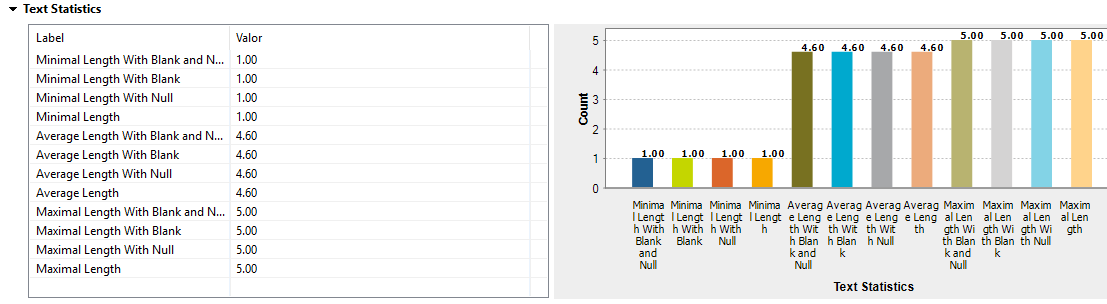
A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo FECNAC.

Con este análisis se obtiene como resultado que existen 3306 campos vacíos, lo que representa un 13.22% del total, además se muestra como existe bastante diferencia entre los formatos de fecha ingresados.

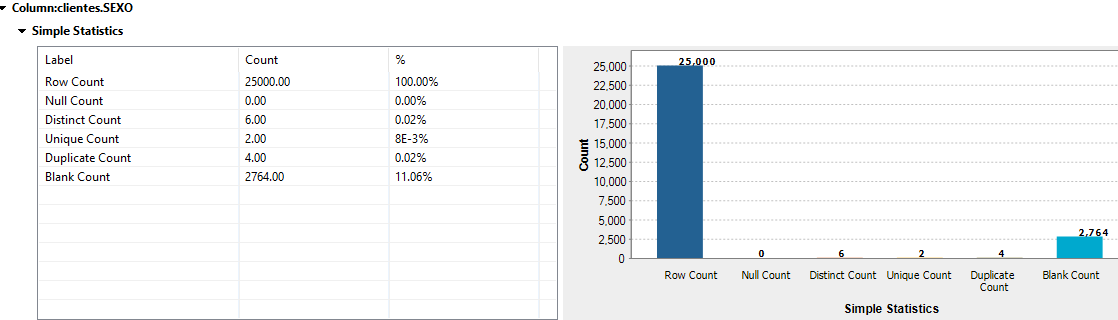
**Estadísticas resumen**

**Estadísticas Simples para el atributo IDCLIENTE**

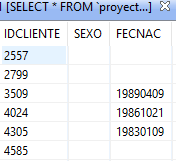
Del total de 25000 atributos, se hace un conteo sobre su totalidad, donde todos los campos son únicos y distintos.

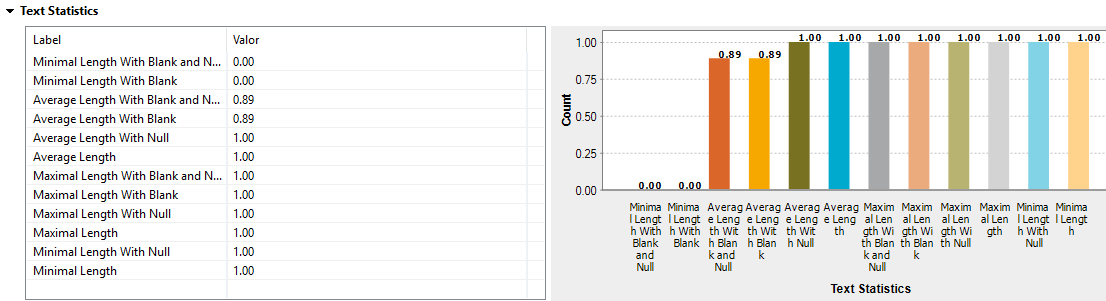
**Estadísticas textuales para el atributo IDCLIENTE**

Se muestra un promedio de valores mínimos de un 1.00, valores promedio de 4.60 y valores máximos de 5.00.

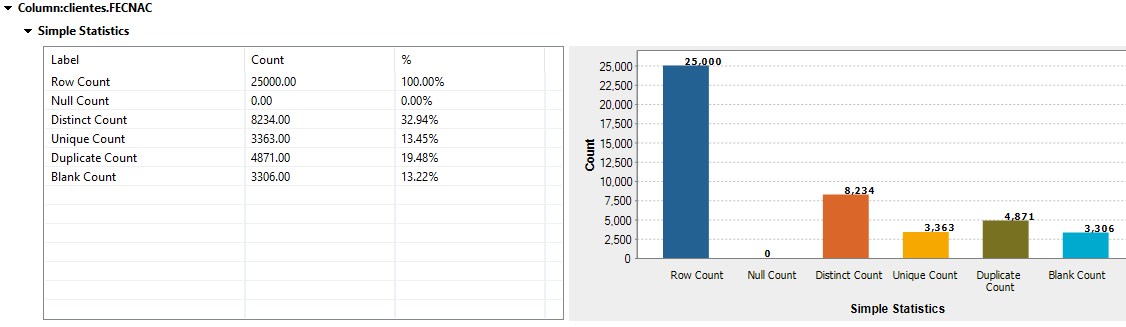
**Estadísticas Simples para el atributo SEXO**

Se muestran 0.00% campos nulos, 0.02% distintos, 0.02% duplicados y 11.06% en blanco.

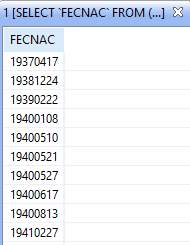
A continuación, se muestran algunos de los valores duplicados.

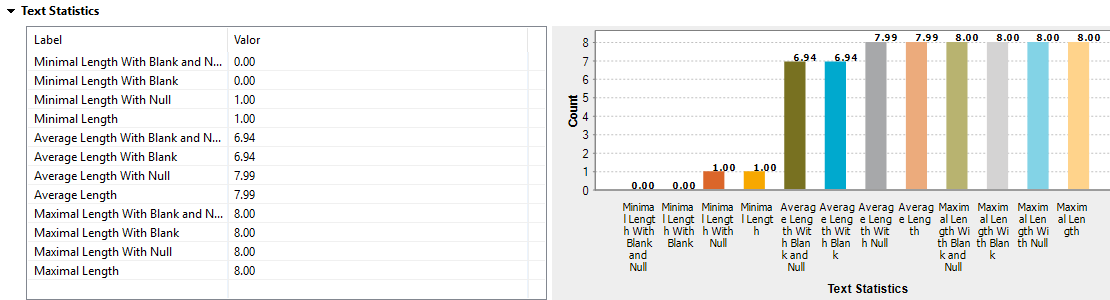
**Estadísticas textuales para el atributo SEXO**

Se muestran valores en blanco mínimos de 0.00, valores nulos mínimos de 0.89, valores promedio y máximos de 1.00.

**Estadísticas Simples para el atributo FECNAC**

Se muestra que no existen campos nulos, un total de 32.94% de campos distintos, 13.45 de campos únicos, 19.48% de campos duplicados y un 13.22% de campos en blanco.

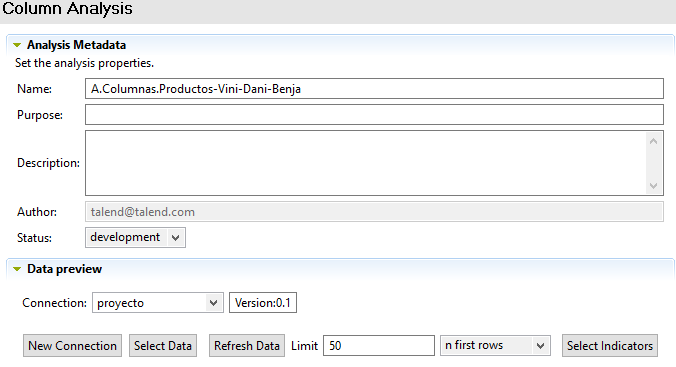
A continuación, se muestran algunos valores de los campos duplicados.

**Estadísticas textuales para el atributo FECNAC**

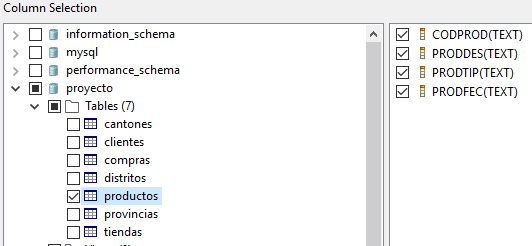
Se muestra como los mínimos en blanco son de 0.00, los mínimos nulos de 1.00, el promedio en blanco de 6.94, el promedio de nulos de 7.99 y los máximos de 8.00.

## **Tabla Productos**

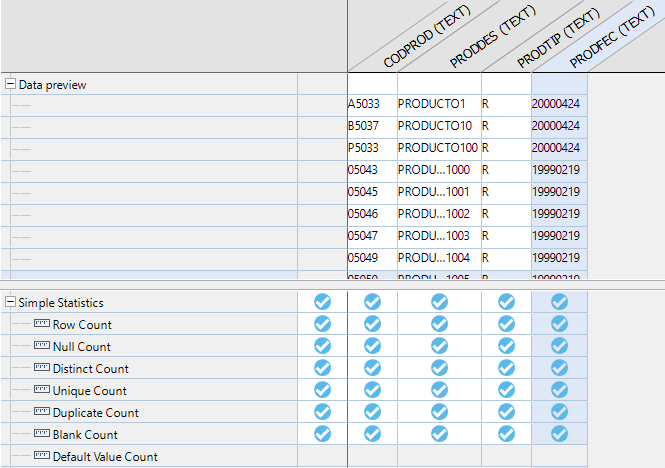
**Análisis Básico de columna**

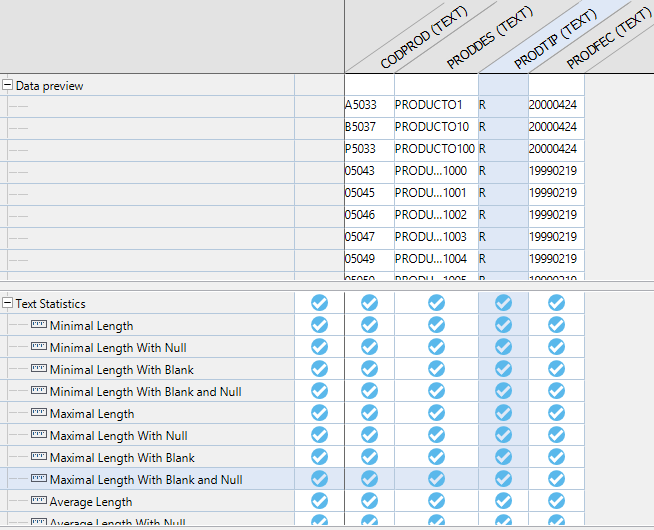
Se crea el análisis bajo el nombre A.Columnas.Productos-Vini-Dani-Benja

Se selecciona los indicadores de la tabla



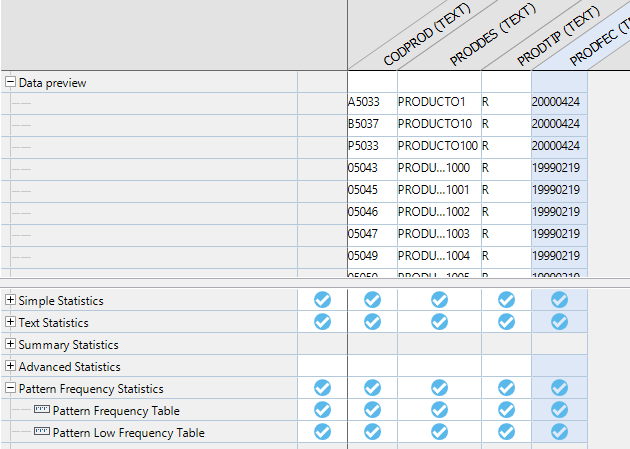
Se seleccionan los indicadores y se elige “Estadísticas Simples” para todos los atributos.

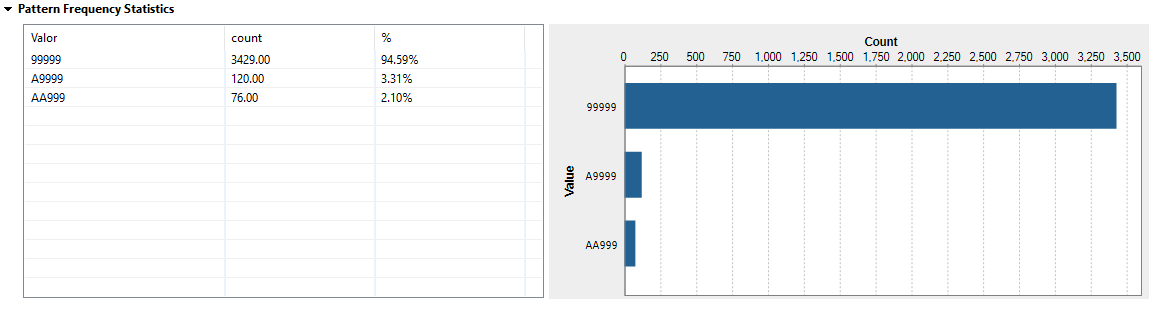


Se selecciona “Estadísticas de Texto” para este tipo de atributos.

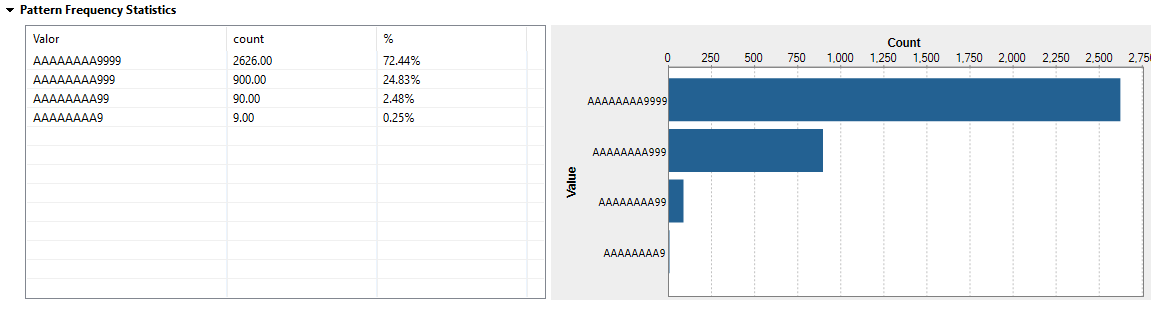
**Análisis de Frecuencia de Patrones**

Se realiza el análisis de frecuencia de patrones para los atributos de la tabla Productos.

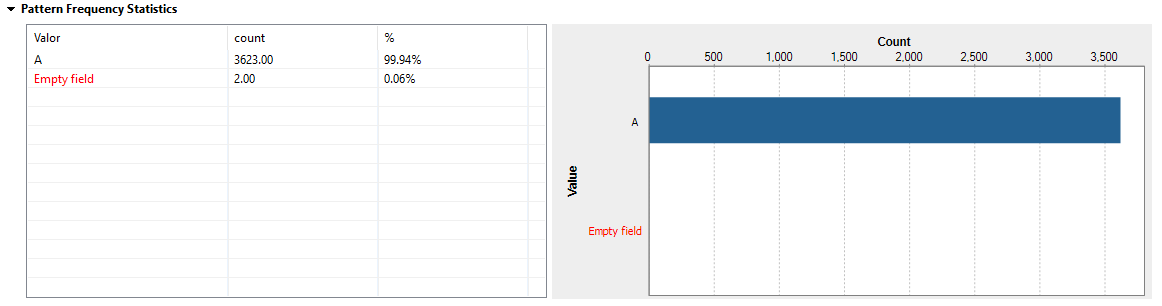


A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo CODPROV.

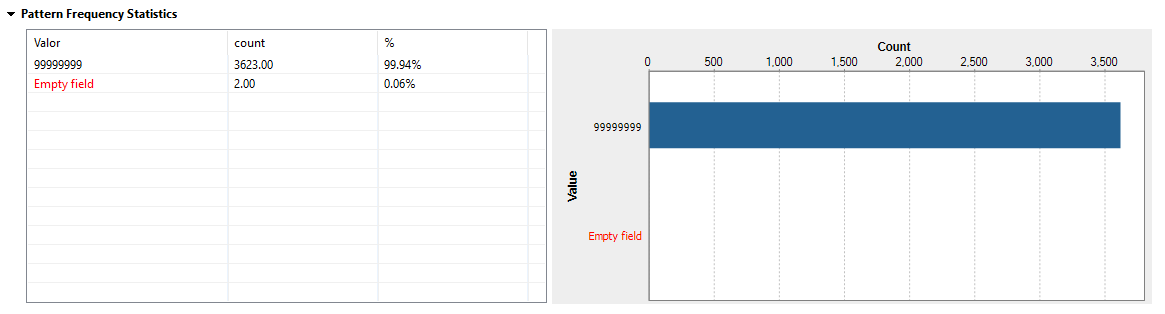
En este se muestran 3 formatos distintos, en el cual el 94.59% representa valores numéricos de 6 dígitos, los otros dos corresponden a valores alfanuméricos. Todos estos son valores válidos.

A continuación se presenta el resultado del análisis de frecuencia de patrones para el atributo PRODDES.

Esta muestra valores de formatos para las descripciones, en el cual todos resultan válidos.

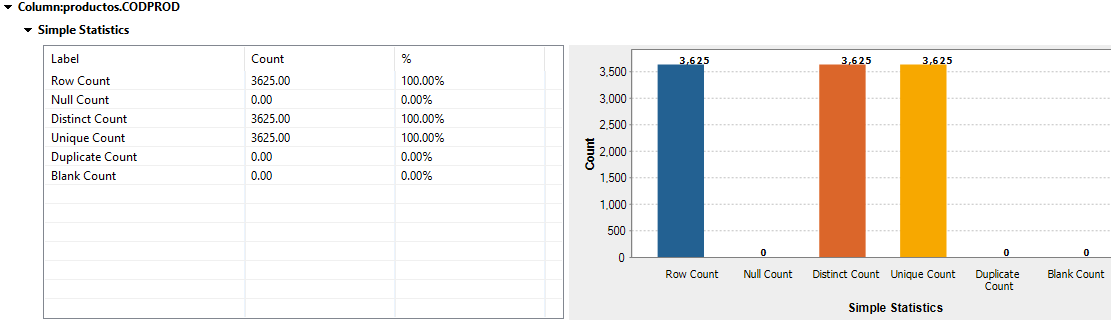
A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo PRODTIP.

Se muestra que existe un 0.06% de productos sin este atributo, el 99.94% representa el valor textual válido correspondiente a “R” y “B”.

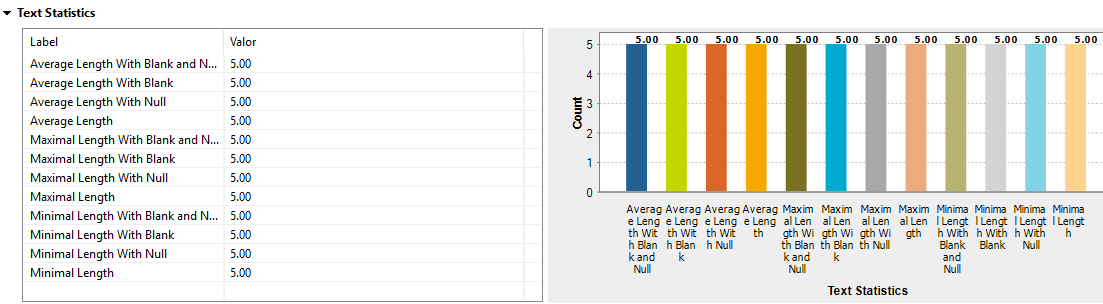
A continuación, se presenta el resultado del análisis de frecuencia de patrones para el atributo PRODFEC.

El resultado anterior muestra un conteo de 2.00 campos vacíos lo cual representa un total de 0.06%, el 99.94% de los restantes poseen el mismo formato.

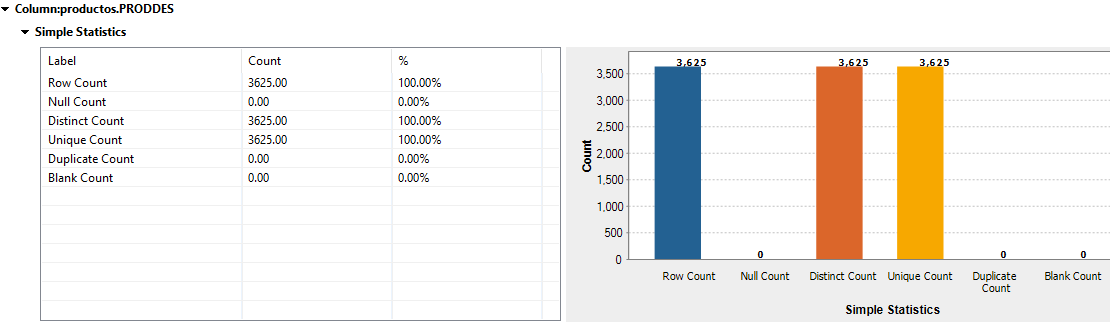
**Estadísticas resumen**

**Estadísticas Simples para el atributo CODPROV**

El análisis se realizó en los 3625 atributos de la columna CODPROD, todos los valores son únicos y distintos.

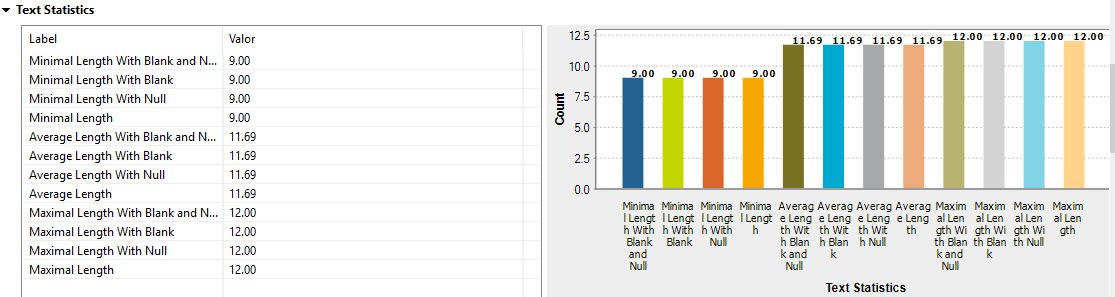
**Estadísticas textuales para el atributo CODPROV**

Los resultados muestran valores promedio, mínimos y máximos de 5.00.

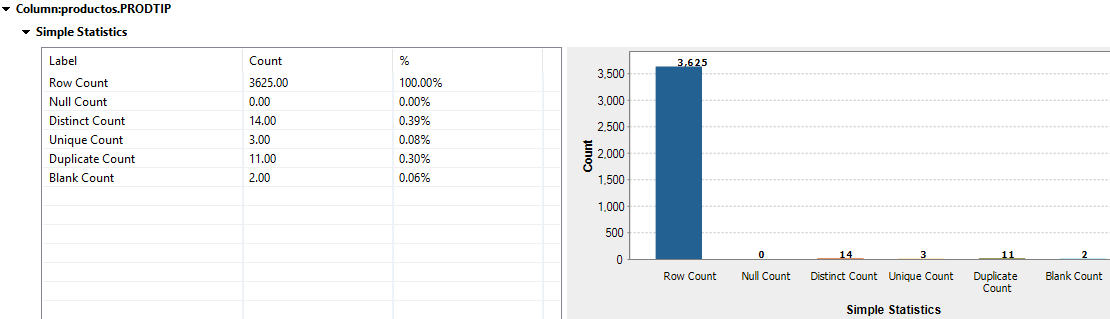
**Estadísticas Simples para el atributo PRODDES**

Todos los campos presentes son únicos y distintos.

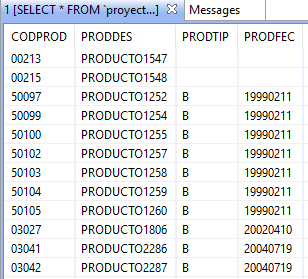
**Estadísticas textuales para el atributo PRODDES**

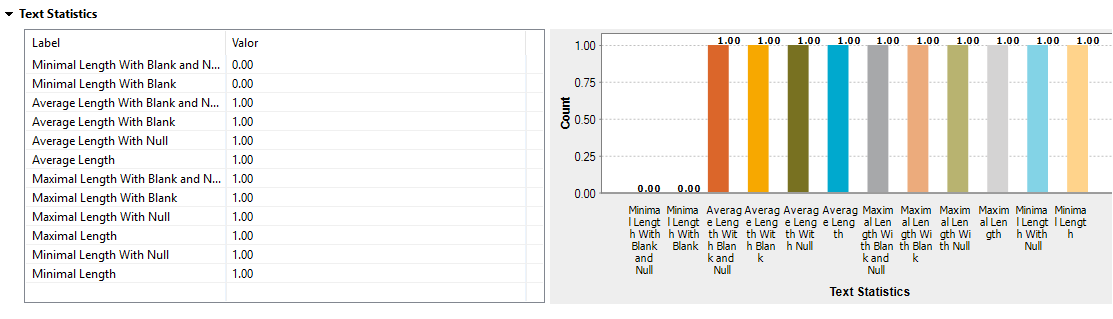


Se muestra un promedio de 9.00 para los valores mínimos, un 11.69 para los valores promedio y un 12.00 para los máximos.

**Estadísticas Simples para el atributo PRODTIP**

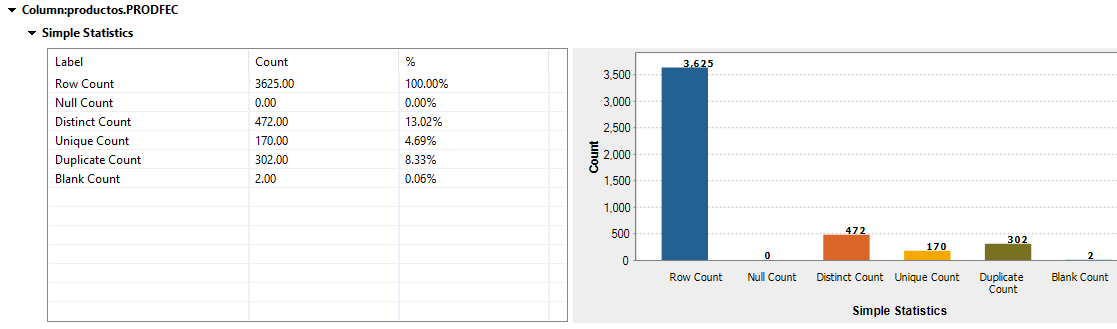
Este análisis muestra un total de 0 valores nulos, 14 distintos, 3 únicos, 11 duplicados y 2 en blanco.

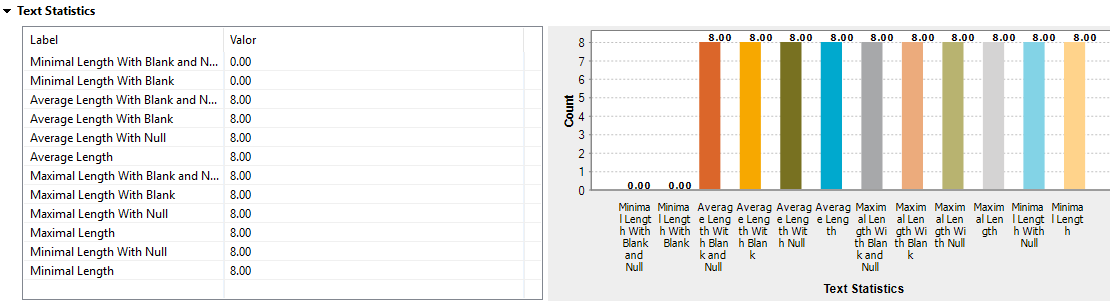
A continuación, se muestran los valores duplicados.

**Estadísticas textuales para el atributo PRODTIP**

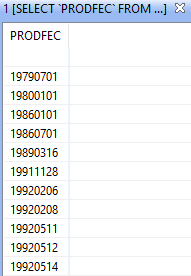
Se muestra un mínimo de valores en blanco de 0.00, un mínimo de nulos de 1.00, un promedio y un máximo de valores de 1.00.

**Estadísticas Simples para el atributo PRODFEC**

El resultado muestra un total de 0 valores nulos, 472 distintos, 170 únicos, 302 duplicados y 2 en blanco.

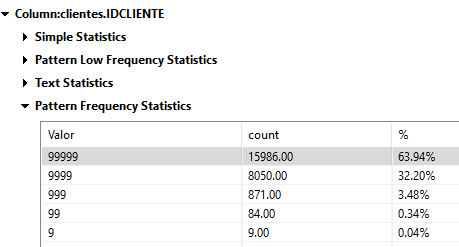
**Estadísticas textuales para el atributo PRODFEC**

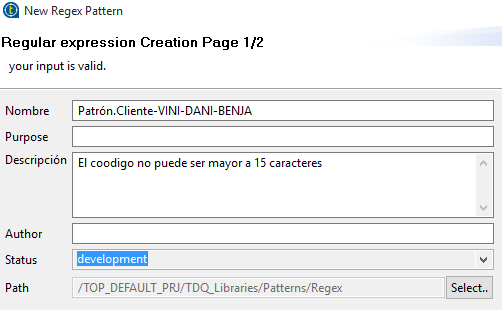
Se muestra un mínimo de valores en nulos de 0.00, un mínimo de valores en blanco de 8.00, un promedio y un máximo de valores de 8.00.

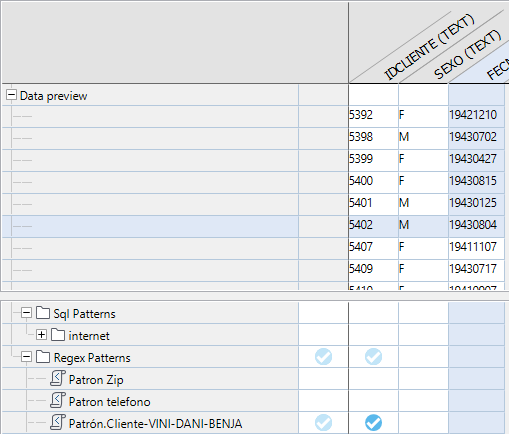
A continuación, se muestran algunos de los valores duplicados.

# **Agregación de Patrones**

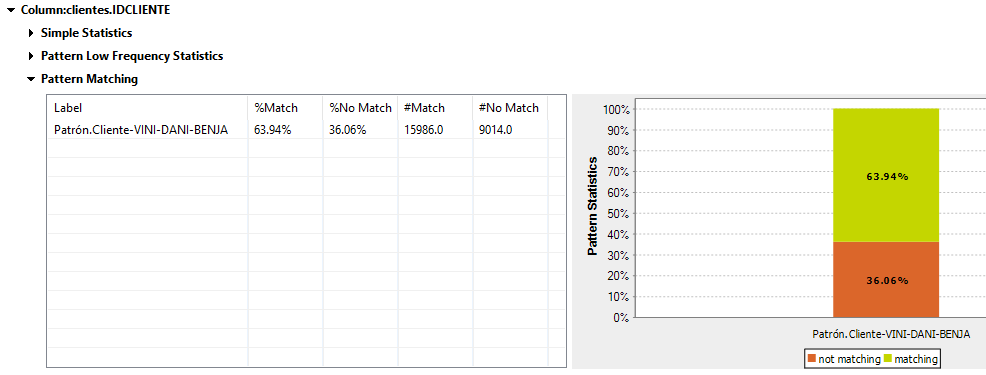
## **IDCLIENTE**

En Analysis Results, buscar los resultados de Pattern Frequency Statistics para el campo IDCLIENTE, dar click derecho y elegir Generate Regex Pattern:

Se crea el siguiente patrón:

En la sección Analysis Settings se selecciona el patrón creado para el atributo correspondiente:

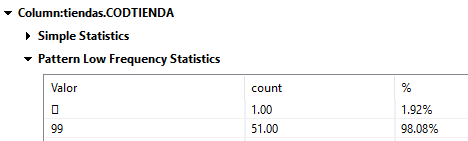
El resultado de la aplicación de este patrón es la siguiente:



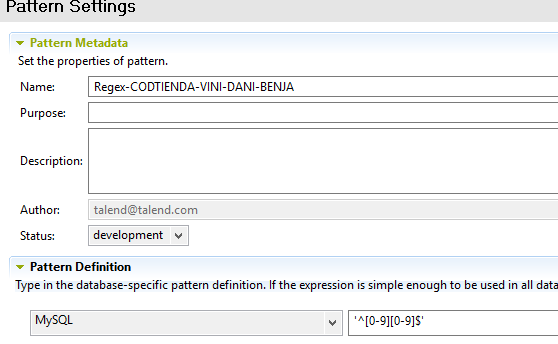
Como se seleccionó un patrón para el valor más alto que tenía 5 caracteres, el resultado para este es de 63.34% de valores que lo cumplen

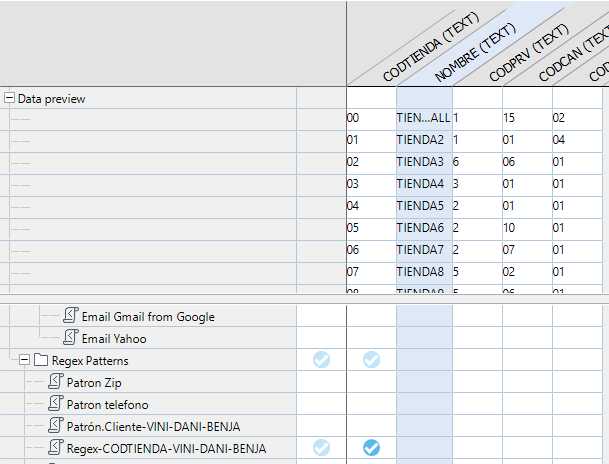
## **CODTIENDA**

En Analysis Results, buscar los resultados de Pattern Frequency Statistics para el campo CODTIENDA, dar click derecho y elegir Generate Regex Pattern. Este patrón solicita que sean 2 caracteres.

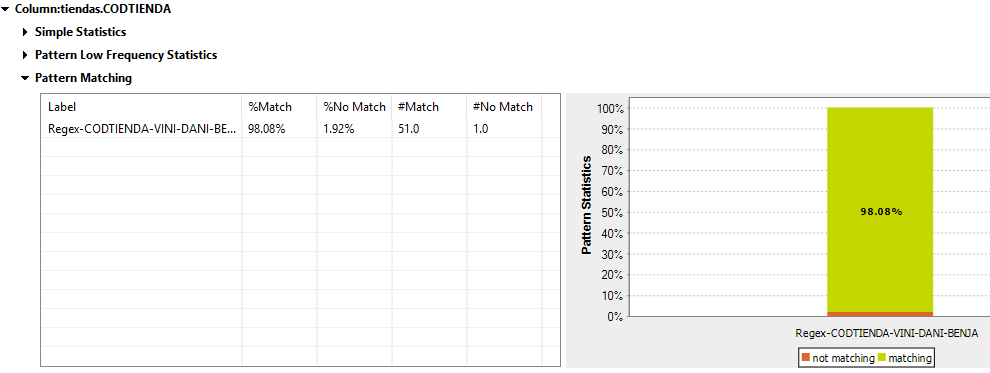


Se agrega el siguiente patrón:



En la sección Analysis Settings se selecciona el patrón creado para el atributo correspondiente

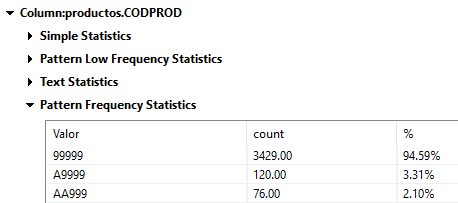
El resultado es el siguiente:



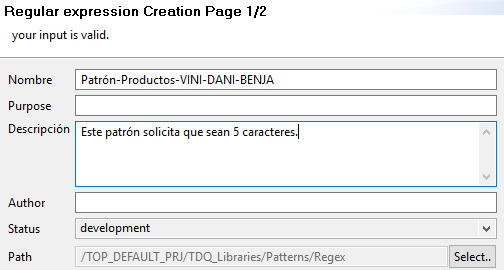
El resultado muestra un 98.08% de calores que cumplen en patrón.

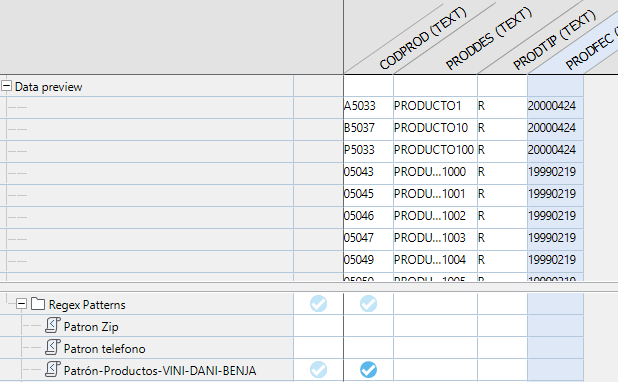
## **CODPROD**

Para este patrón se solicita que sean 5 caracteres. En Analysis Results, buscar los resultados de Pattern Frequency Statistics para el campo CODPROD, dar click derecho y elegir Generate Regex Pattern

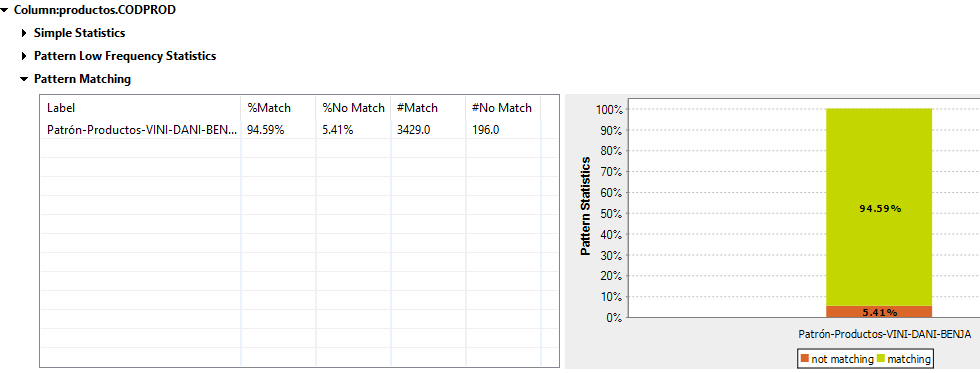


Se agrega el siguiente patrón:



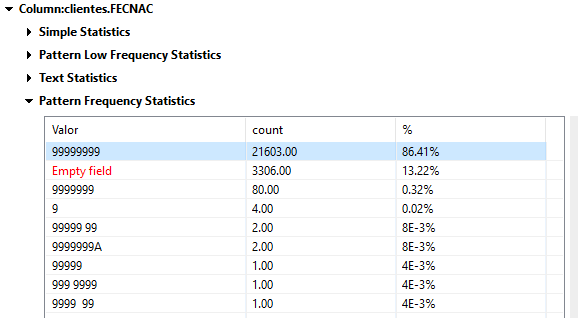
En la sección Analysis Settings se selecciona el patrón creado para el atributo correspondiente

El resultado es el siguiente:

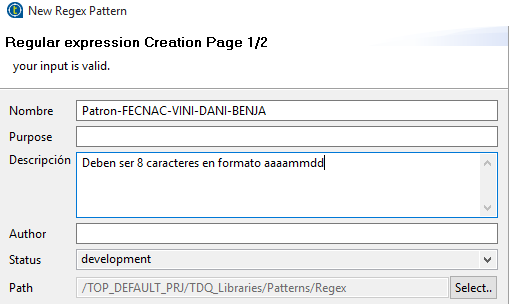
Se muestra como un 94.59% cumple de los resultados cumplen con este patrón.

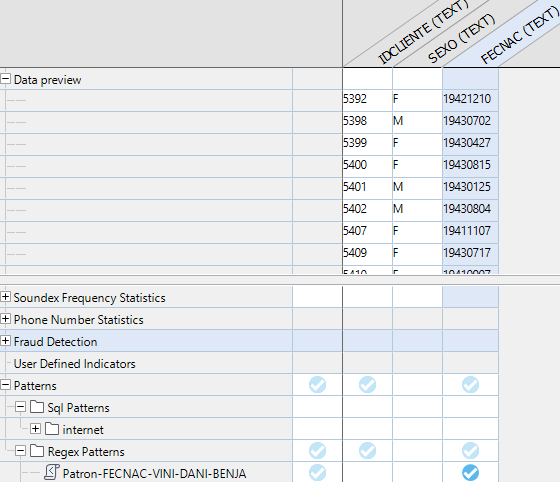
## **FECNAC**

Para este patrón se solicita que sean 8 caracteres. En Analysis Results, buscar los resultados de Pattern Frequency Statistics para el campo FECNAC, dar click derecho y elegir Generate Regex Pattern

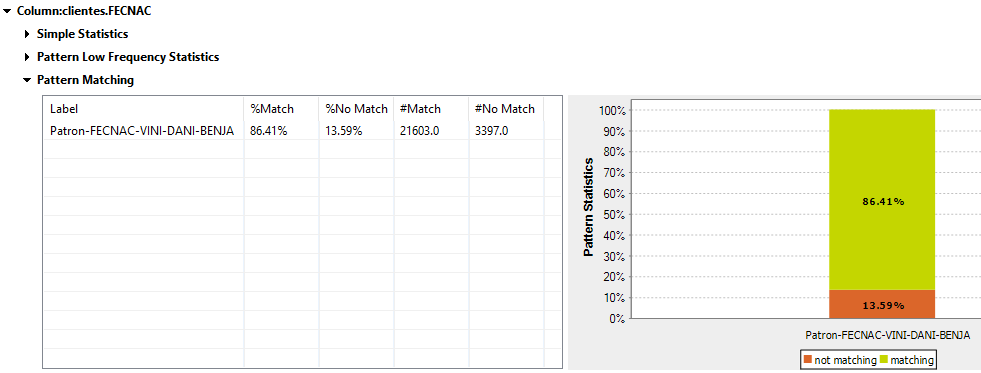


Se crea el siguiente patrón:



En la sección Analysis Settings se selecciona el patrón creado para el atributo correspondiente:

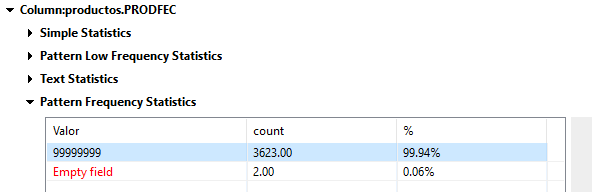
El resultado es el siguiente:



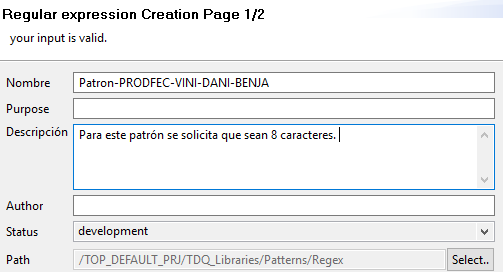
Se muestra un 86.41% de resultados que cumplen el patrón.

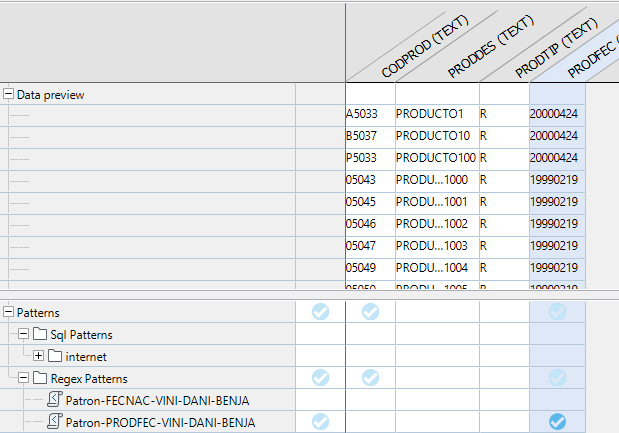
## **PRODFEC**

Para este patrón se solicita que sean 8 caracteres. En Analysis Results, buscar los resultados de Pattern Frequency Statistics para el campo PRODFEC, dar click derecho y elegir Generate Regex Pattern

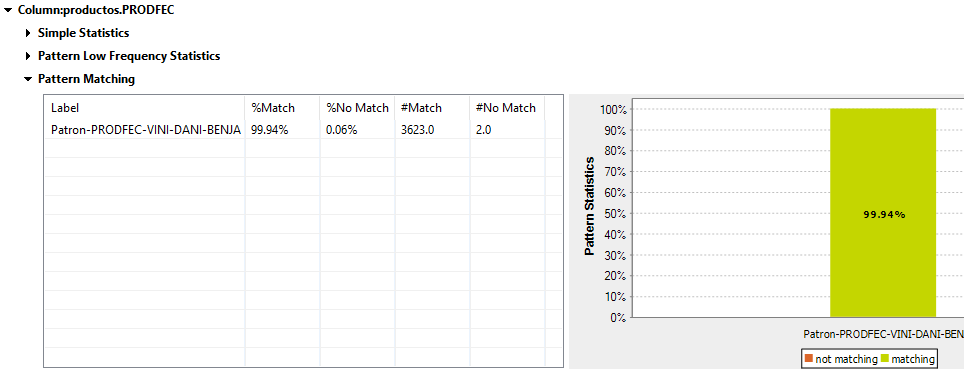


Se crea el siguiente patrón:



En la sección Analysis Settings se selecciona el patrón creado para el atributo correspondiente:

El resultado es el siguiente:

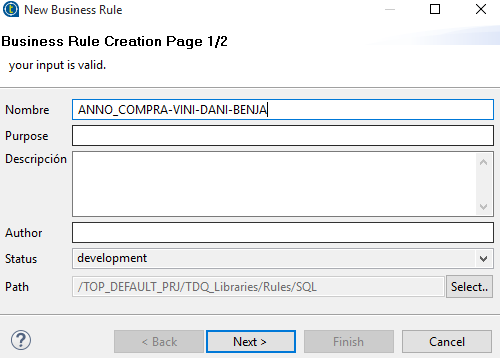


El resultado muestra un 99.94% de valores que cumplen el patrón.

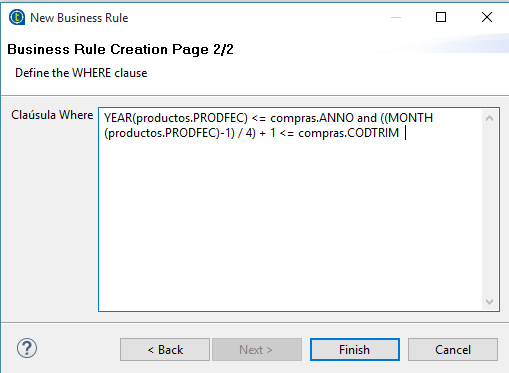
# **Reglas de Negocio**

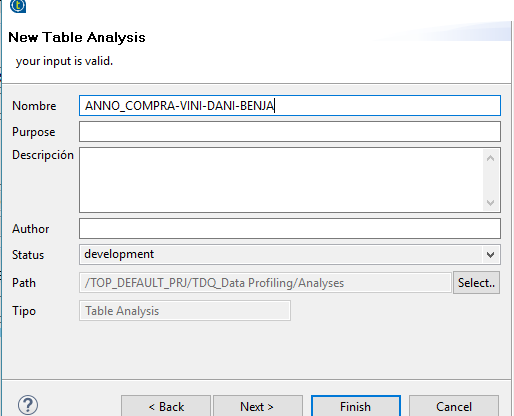
## **Regla fecha de compra**

Se debe crear, implementar y ejecutar una regla de negocio para validar que el año en que un cliente adquirió un producto sea igual o mayor al año en que el producto se empezó a vender.

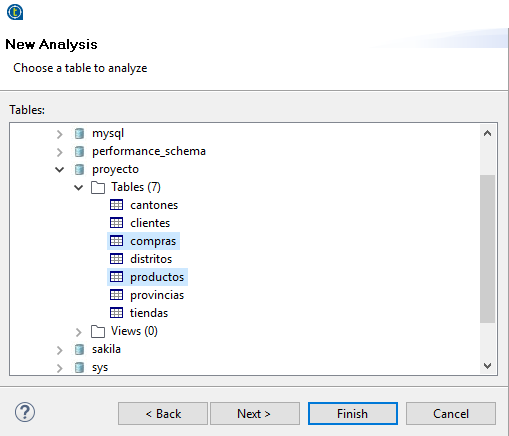
Para esto se crea la regla de negocio con el nombre ANNO\_COMPRA-VINI-DANI-BENJA.

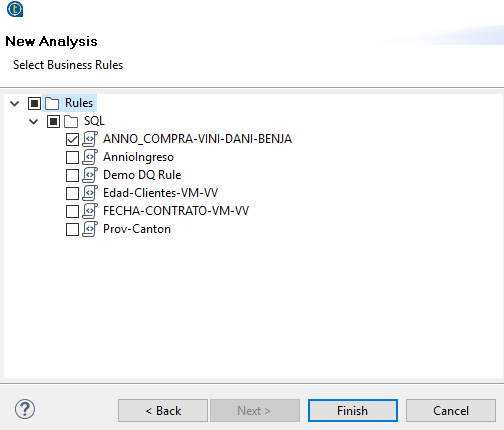
Se crea la regla de negocio que verifica que la compra del cliente se de en una fecha igual o posterior a la fecha en que el producto comienza a comercializarse.



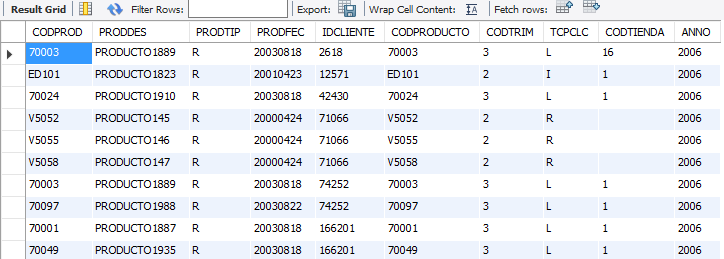
Se crea el análisis bajo el mismo nombre de la regla.

Se seleccionan las tablas bajo las que se llevará a cabo el análisis.



Se selecciona la regla previamente creada.

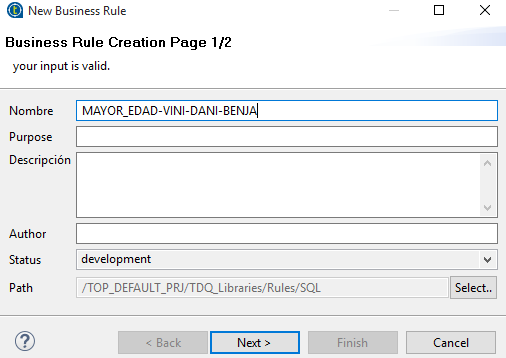
Debido a que existe una diferencia entre los formatos y valores válidos del atributo de código de los productos en ambas tablas, solo existen 4953 compras válidas que se pueden relacionar entre la tabla productos y la tabla compras.

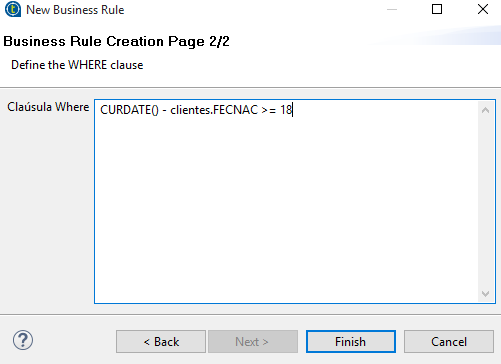
El resultado de esta regla en MySQL muestra los siguientes resultados.

Esta obtiene como resultado 1814 registros que cumplen dicha regla.

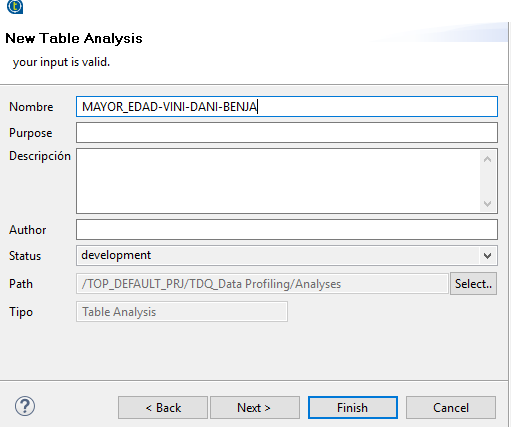
## **Regla Mayoría de Edad**

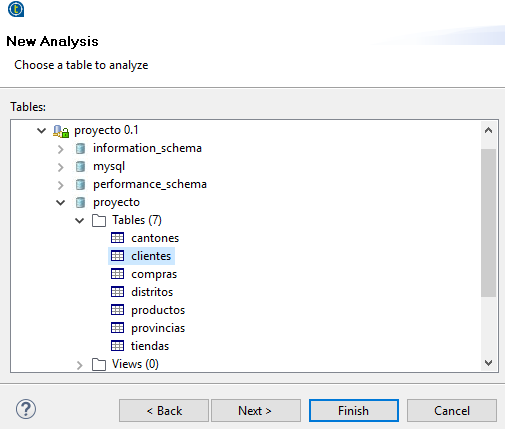
Se debe crear, implementar y ejecutar una regla de negocio para validar que el año en que un cliente adquirió un producto sea igual o mayor al año en que el producto se empezó a vender.

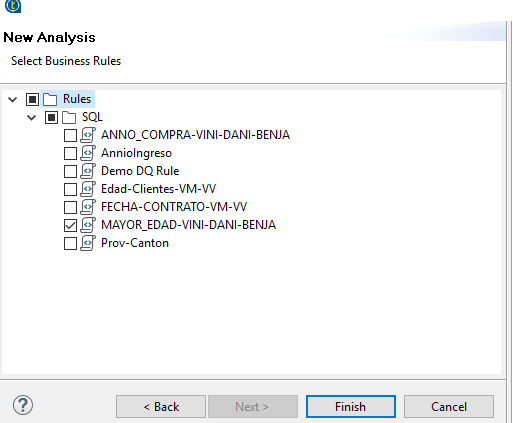
Se crea la regla MAYOR\_EDAD-VINI-DANI-BENJA.

La cláusula Where para esta regla toma la fecha actual y le resta la fecha de nacimiento del cliente, los resultados que devuelve son los clientes mayores de 18 años.

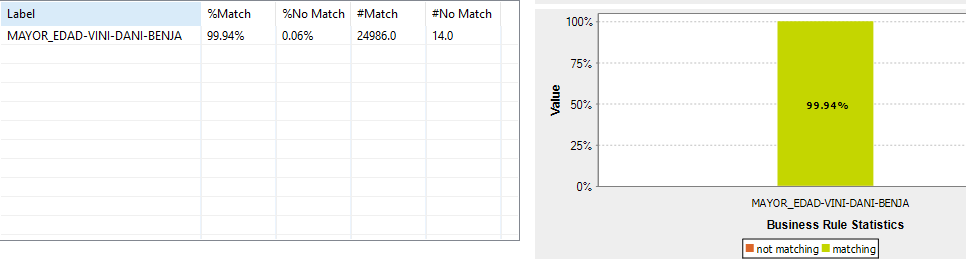
Se crea un nuevo análisis bajo el nombre de la regla.

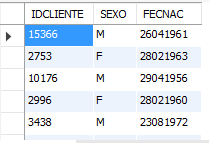


Se selecciona la base de datos proyecto y la tabla clientes.

Se asigna la regla correspondiente.

El resultado de la implementación de esta regla muestra lo siguiente.

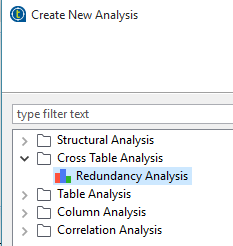
El 99.94% de los registros cumplen la regla de ser mayores de edad.

Al comprobar esta regla en MySQL para obtener los clientes menores de edad se obtiene lo siguiente.

Este resultado muestra que dichos clientes incumplen el estándar de aaaammdd, lo cual hace que al tener su fecha de nacimiento en ddmmaaaa, aparezcan como menores de edad.

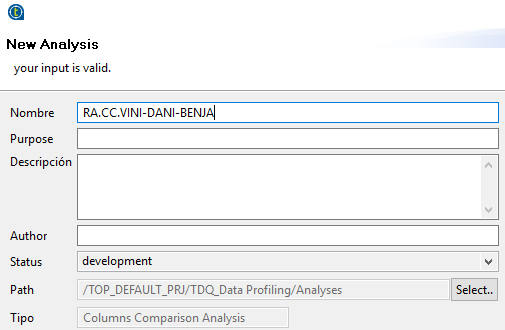
# Perfilado de Relaciones.

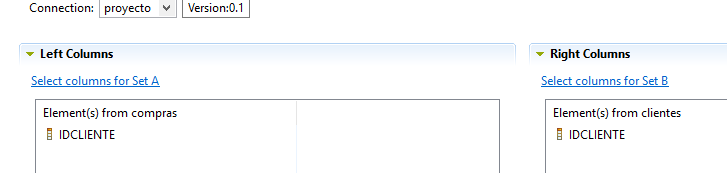
En Data Profiling se crea un nuevo análisis, se escoge Cross Table Analysis, luego Redundancy Analysis.

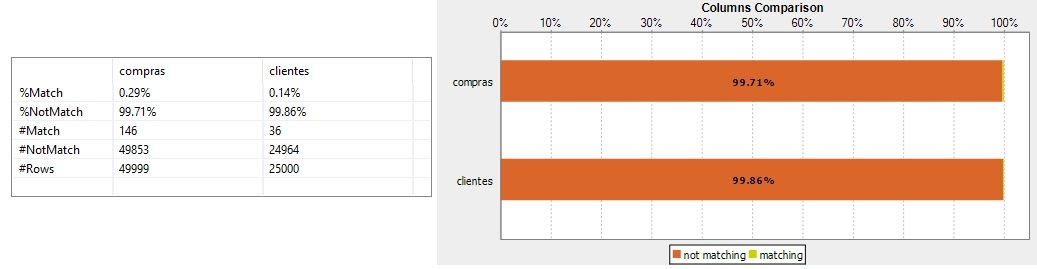


## **COMPRAS-CLIENTES**

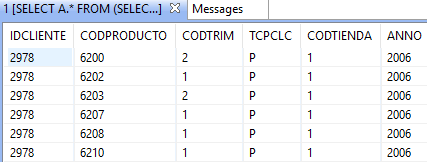
Se crea el siguiente análisis:

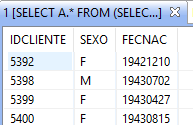


En la sección Select columns for Set A, se selecciona la tabla compras y el atributo IDCLIENTE, en el Set B la tabla clientes y el atributo IDCLIENTE.

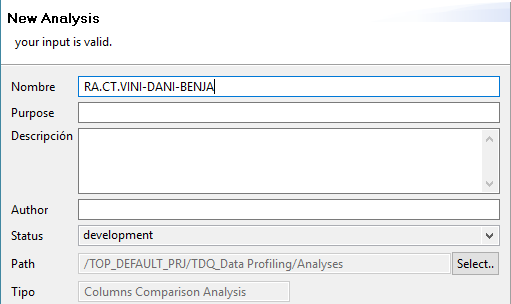
Los resultados son los siguientes:

El resultado muestra un not match de 49853 en la tabla compras y 24964 en la tabla clientes. Dichos porcentajes corresponden a un 99.71% y 99.86% respectivamente.

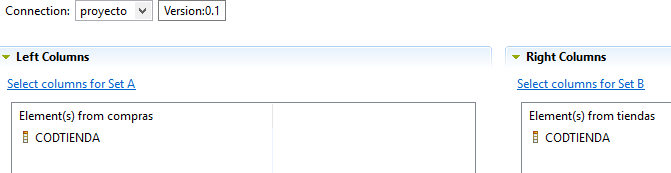
En la tabla compras estos son algunos de dichos resultados:

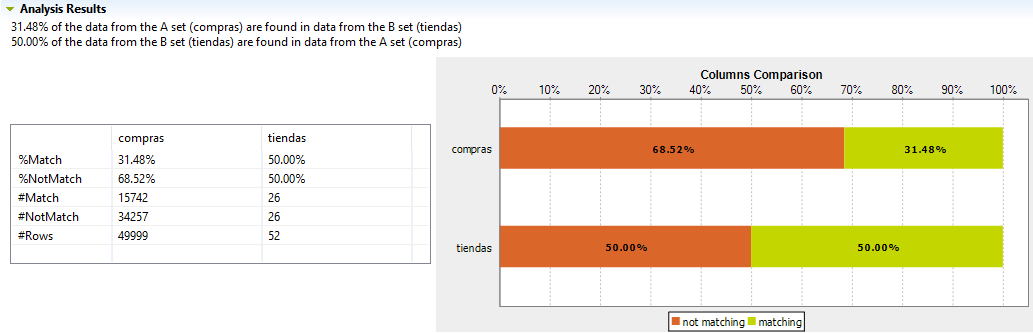
En la tabla clientes estos son algunos de los resultados:

## **COMPRAS-TIENDAS**

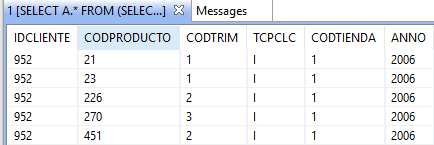
Se crea el siguiente análisis:

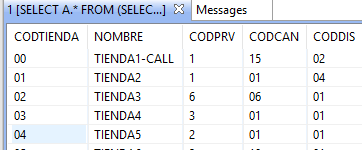
En la sección Select columns for Set A, se selecciona la tabla compras y el atributo CODTIENDA, en el Set B la tabla tiendas y el atributo CODTIENDA.



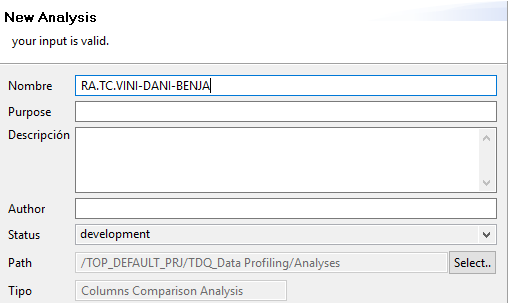
El resultado es el siguiente:

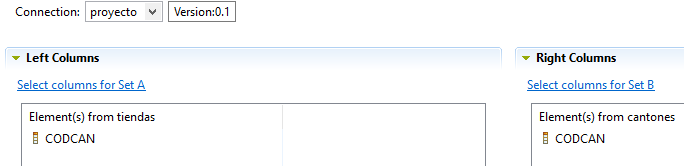
En compras existe un 31.48% de valores encontrados en las tiendas, en las tiendas existe un 50% de valores encontrados en las compras.

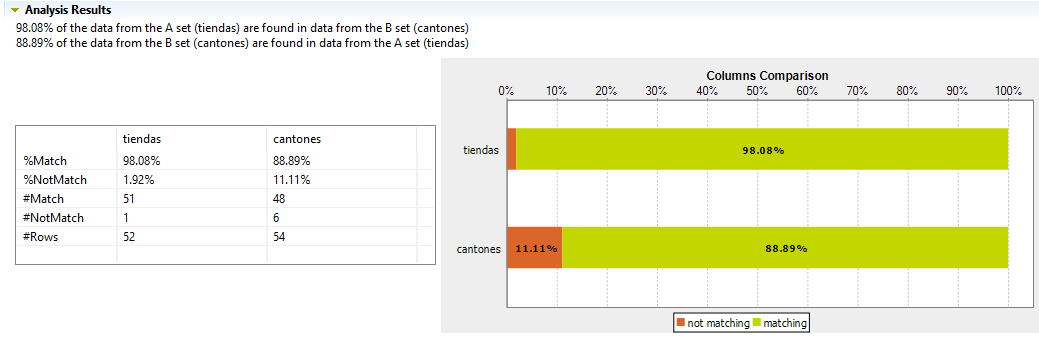
Algunos de los valores no encontrados en la tabla compra son los siguientes:

Algunos resultados no encontrados en la tabla tienda son los siguientes:

## **TIENDAS-CANTON**

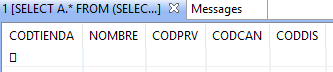
Se crea el siguiente análisis:

En la sección Select columns for Set A, se selecciona la tabla tiendas y el atributo CODCAN, en el Set B la tabla cantones y el atributo CODCAN.

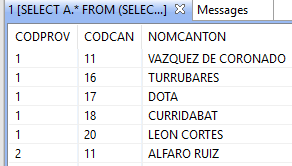
Los resultados son los siguientes:

En la tabla tiendas existe un 98.08% de resultados que están en los cantones, en los cantones existe un 88.89% de valores encontrados en las tiendas.

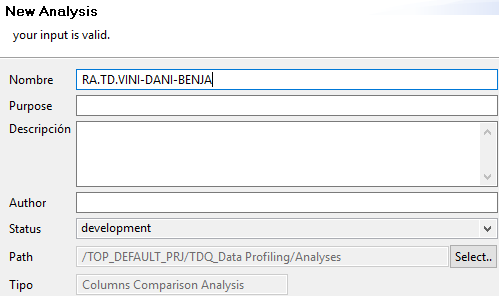
El valor no encontrado en la tabla tiendas es el siguiente:

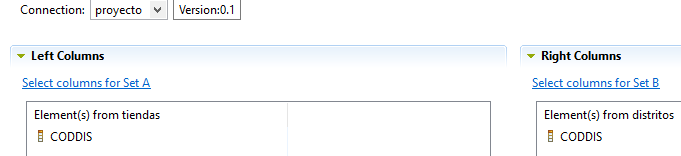


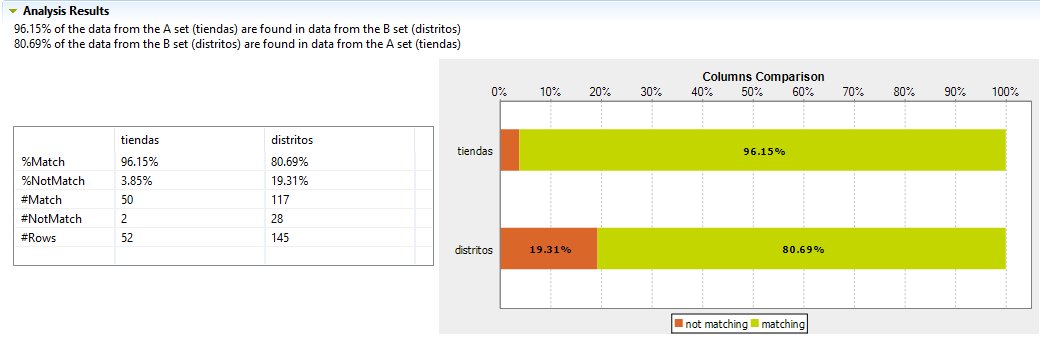
Dicho valor corresponde a un registro vacío.

En la tabla cantones existen los siguientes 6 valores no encontrados en tiendas:

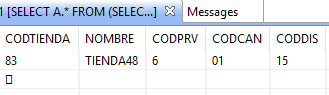
## **TIENDAS-DISTRITO**

Se crea el siguiente análisis:

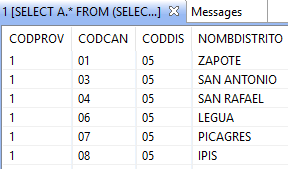
En la sección Select columns for Set A, se selecciona la tabla tiendas y el atributo CODDIS, en el Set B la tabla distritos y el atributo CODDIS.

El resultado es el siguiente:

Se muestra que la tabla tiendas tiene un 96.15% de resultados presentes en los distritos, la tabla distritos tiene un 80.69% de registros presentes en tiendas.

Los 2 resultados de tiendas no presentes en distritos son los siguientes:

Algunos de los registros en la tabla distritos no presentes en tiendas son los siguientes:



Catálogo de reglas de calidad de datos

En esta sección se describen las reglas de calidad asociadas a las tablas del proyecto. Las cuales son diseñadas para evaluar los atributos de dichas tablas.

A continuación, se detallan las partes del listado de reglas:

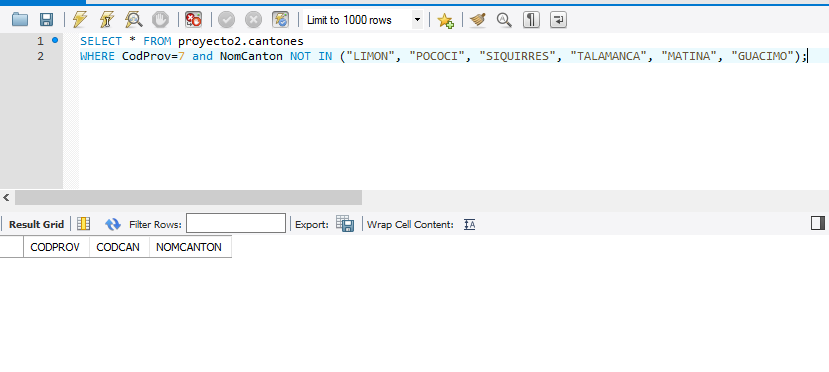
|  |  |
| --- | --- |
| Atributo | Descripción |
| Identificador de la regla | Identificador único para cada regla |
| Nombre de la regla | Nombre descriptivo de la regla |
| Definición formal | Definición técnica |
| Definición informal | Descripcipon en lenguaje natural |
| SQL | Script SQL, para validar la regla |

# Restricciones de valores válidos

## Tabla: Cantones

|  |
| --- |
| Regla #1  Valores válidos para cantones de limón  Definición formal  Cantones.CodProv= 7, Cantones.NomCanton IN ("LIMON", "POCOCI", "SIQUIRRES", "TALAMANCA", "MATINA", "GUACIMO")  Definición informal  De la provincia de Limón, la cual su identificador es el 7, los valores permitidos para el nombre del Canton son Limon, Pococi, Siquirres, Talamanca, Matina o Guacimo.  SQL  SELECT \* FROM proyecto2.cantones  WHERE CodProv=7 and NomCanton NOT IN ("LIMON", "POCOCI", "SIQUIRRES", "TALAMANCA", "MATINA", "GUACIMO"); |

### Screenshot

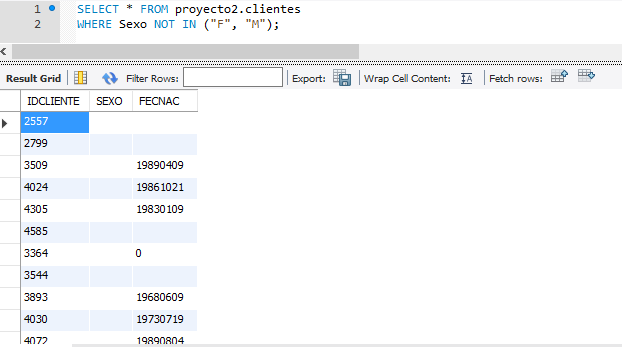


La consulta no devuelve ningún resultado, debido a que la provincia de Limón tiene correctos los nombres de sus cantones.

## Tabla: Clientes

|  |
| --- |
| Regla #2  Valores válidos sexo de cliente  Definición formal  Clientes.Sexo IN (“F” O “M”)  Definición informal  Los valores permitidos para el sexo del cliente son “F” que significa femenino o “M” que significa masculino  SQL  SELECT \* FROM proyecto2.clientes  WHERE Sexo NOT IN ("F", "M"); |

### Screenshot

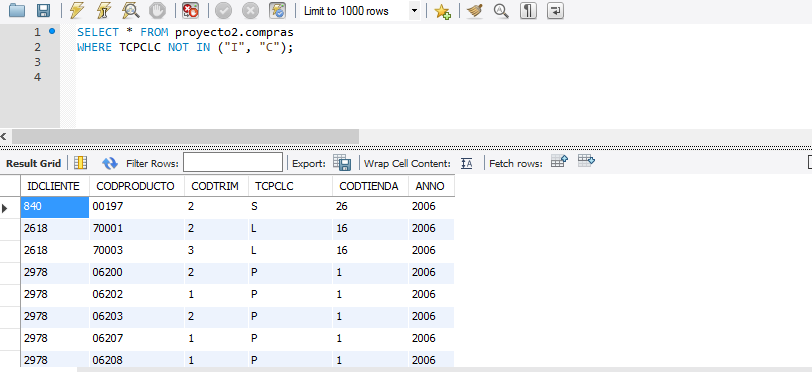


Se muestran los resultados de registro que no tienen correctamente definido el sexo como F o M.

## Tabla: Compras

|  |
| --- |
| Regla #3  Valores válidos tipos de cliente  Definición formal  Compras.TCPCLC IN (“I” O “C”)  Definición informal  Los valores permitidos para el tipo de cliente son “I” que significa individual o “C” que significa corporativo  SQL  SELECT \* FROM proyecto2.compras  WHERE TCPCLC NOT IN ("I", "C"); |

### Screenshot

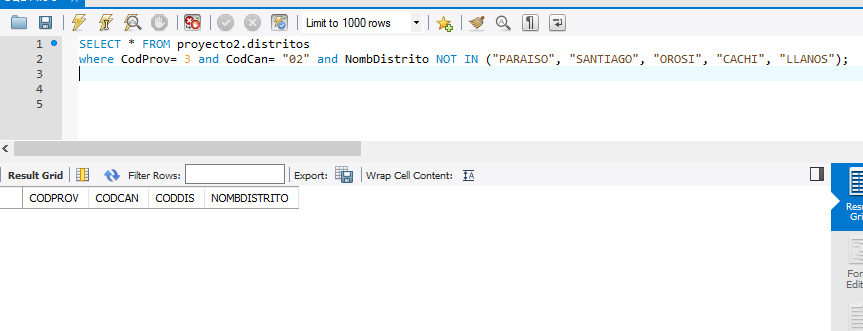


Se muestran los registros cuyo tipo de compra no es “I” o “C”

## Tabla: Distritos

|  |
| --- |
| Regla #4  Valores válidos para de distrito para el cantón de Paraiso de Cartago  Definición formal  Distritos.CodProv= 3 and Distritos.CodCan=2, Distritos.NombDistrito IN ("PARAISO", "SANTIAGO", "OROSI", "CACHI", "LLANOS")  Definición informal  De la provincia de Cartago, cuyo identificador es el 3, cantón de Paraiso, cuyo identificador es 2, los distritos permitidos son Paraiso, Santiago, Orosi, Cachi y Llanos.  SQL  SELECT \* FROM proyecto2.distritos  where CodProv= 3 and CodCan= "02" and NombDistrito NOT IN ("PARAISO", "SANTIAGO", "OROSI", "CACHI", "LLANOS"); |

### Screenshot

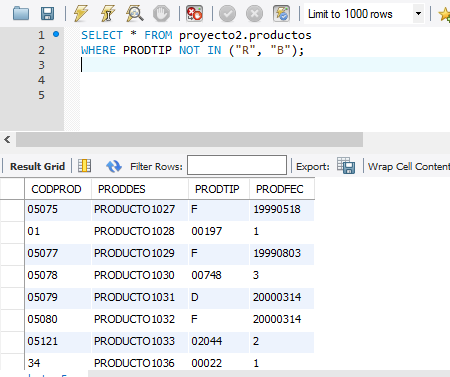


La consulta no devuelve ningún resultado, debido a que el cantón de Paraíso de Cartago tiene correctos los nombres de sus distritos.

## Tabla: Productos

|  |
| --- |
| Regla #5  Valores válidos para tipo de producto  Definición formal  Productos.ProdTip IN (“R” O “B”)  Definición informal  Los valores permitidos para el tipo de producto son “R” o “B”  SQL  SELECT \* FROM proyecto2.productos  WHERE PRODTIP NOT IN ("R", "B"); |

### Screenshot

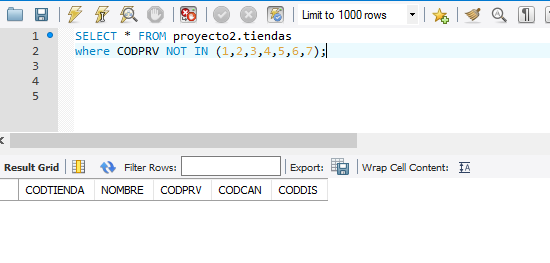


Se muestran los registros de los productos cuyo tipo de producto es diferente de R o B.

## Tabla: TIENDAS

|  |
| --- |
| Regla #6  Valores válidos para la provincia de la tienda  Definición formal  Tiendas.CodPrv IN (1,2,3,4,5,6,7)  Definición informal  Los valores permitidos para el código de la provincia de la tienda pueden ser del 1 al 7.  SQL  SELECT \* FROM proyecto2.tiendas  where CODPRV NOT IN (1,2,3,4,5,6,7); |

### Screenshot



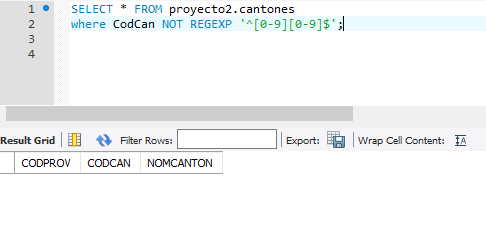
La consulta no devuelve ningún resultado, debido a que todas las tiendas poseen valores de provincia validos

# Restricciones de formato

## Tabla: Cantones

|  |
| --- |
| Regla #7  Formato de código de cantón  Definición formal  Cantones.CodCan LIKE ‘[0-9][0-9]’  Definición informal  El formato del código de cantón es numérico cuyos valores pueden ser del [0-9]. Además posee dos dígitos.  SQL  SELECT \* FROM proyecto2.cantones  where CodCan NOT REGEXP '^[0-9][0-9]$'; |

### Screenshot

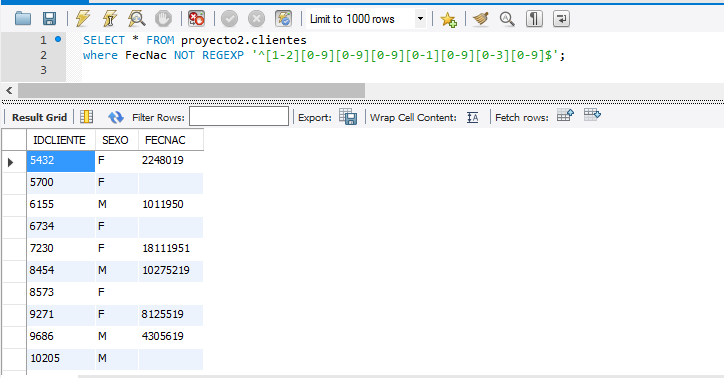


La consulta no devuelve resultados debido a que todos los cantones tienen un código de cantón de 2 dígitos.

## Tabla: Clientes

|  |
| --- |
| Regla #8  Formato de fecha de nacimiento clientes  Definición formal  Clientes.FecNac LIKE '[1-2][0-9][0-9][0-9][0-1][0-9][0-3][0-9]';  Definición informal  La fecha de nacimiento de los clientes debe ser aaaammdd, donde la a representa el año, la m el mes y la d el día. Cabe resaltar que posee 8 dígitos y todos se representan de manera numérica.  SQL  SELECT \* FROM proyecto2.clientes  where FecNac NOT REGEXP '^[1-2][0-9][0-9][0-9][0-1][0-9][0-3][0-9]$'; |

### Screenshot

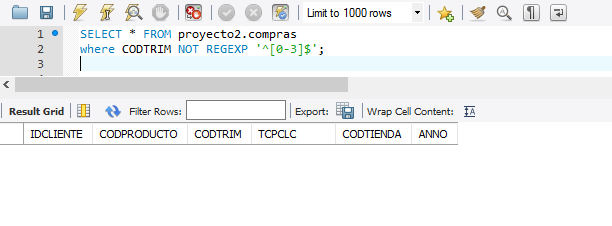


La consulta devuelve la lista de clientes que poseen un formato incorrecto en la fecha de nacimiento.

## Tabla: Compras

|  |
| --- |
| Regla #9  Formato de código de trimestre  Definición formal  Compras. CODPRODUCTO LIKE '[0-3]';  Definición informal  Se define el código del trimestre el cual es de un dígito numéricos y los valores van del 1 al 3.  SQL  SELECT \* FROM proyecto2.compras  where CODTRIM NOT REGEXP '^[0-3]$'; |

### Screenshot

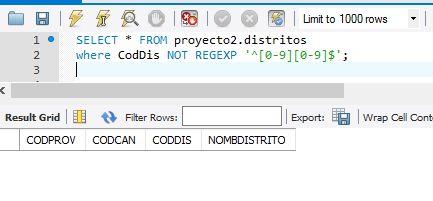


No devuelve ningún resultado debido a que los trimestres de la tabla compras tienen el formato adecuado.

## Tabla: Distritos

|  |
| --- |
| Regla #10  Formato de código de distritos  Definición formal  Distritos.CodDis LIKE ‘[0-9][0-9]’  Definición informal  El formato del código de distrito es numérico cuyos valores pueden ser del [0-9]. Además, posee dos dígitos.  SQL  SELECT \* FROM proyecto2.distritos  where CodDis NOT REGEXP '^[0-9][0-9]$'; |

### Screenshot

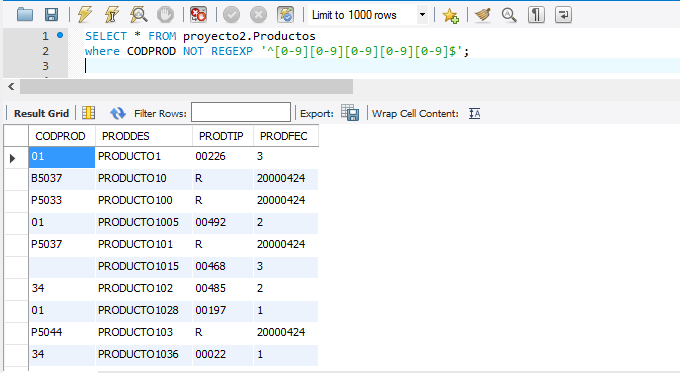


No devuelve ningún resultado debido a que el código de distrito, posee un formato adecuado de dos dígitos.

## Tabla: Productos

|  |
| --- |
| Regla #11  Formato de código de producto  Definición formal  Productos. CODPRODUCTO LIKE '[0-9][0-9][0-9][0-9][0-9]';  Definición informal  Se define el código de producto el cual es de cinco dígitos numéricos.  SQL  SELECT \* FROM proyecto2.Productos  where CODPROD NOT REGEXP '^[0-9][0-9][0-9][0-9][0-9]$'; |

### Screenshot

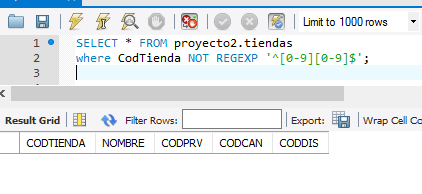


Se muestran los registros cuyo código de producto es diferente de cinco dígitos.

## Tabla: Tiendas

|  |
| --- |
| Regla #12  Formato de código de tienda  Definición formal  Tiendas.CodTienda LIKE '[0-9][0-9]';  Definición informal  Se define el código de producto el cual es de cinco dígitos numéricos.  SQL  SELECT \* FROM proyecto2.tiendas  where CodTienda NOT REGEXP '^[0-9][0-9]$'; |

### Screenshot



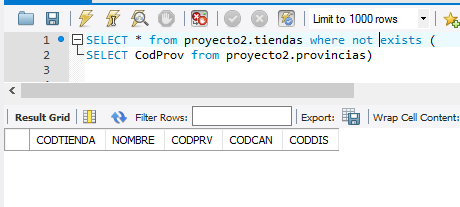
No devuelve ningún resultado debido a que las tiendas tienen correcto el formato del código de la tienda.

# Restricciones de referencia

## Tabla: Tiendas y provincia

|  |
| --- |
| Regla #13  Referencia entre la tabla tienda y provincia  Definición formal  Tiendas. CodPrv EXISTS (Provincias.CodProv);  Definición informal  La tienda debe encontrarse referencia a alguna de las provincias existentes dentro de la tabla provincias, estas van del 1 al 7.  SQL  SELECT \* from proyecto2.tiendas where not exists (  SELECT CodProv from proyecto2.provincias) |

### Screenshot

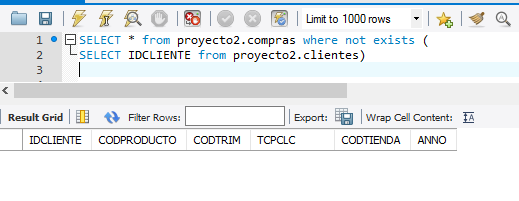


No devuelve ningún registro debido a que todas las tiendas tienen una correcta referencia con la tabla de provincias.

## Tabla: Compras y cliente

|  |
| --- |
| Regla #14  Referencia entre la tabla compras y cliente  Definición formal  Compras. IdCliente EXISTS (Clientes.IdCliente);  Definición informal  La compra debe ser realizada por algún cliente existente dentro de la tabla clientes.  SQL  SELECT \* from proyecto2.compras where not exists (  SELECT IDCLIENTE from proyecto2.clientes) |

### Screenshot



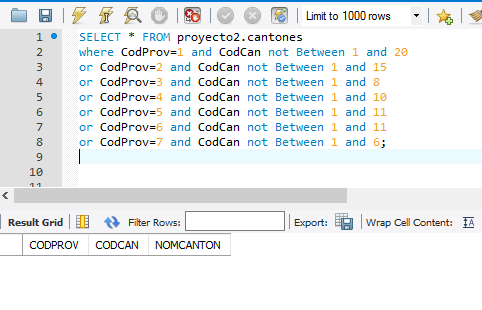
No devuelve ningún registro debido a que todas las compras tienen una correcta referencia con la tabla de clientes.

# Dependencia parcial

## Tabla: Provincia-Cantones

|  |
| --- |
| Regla #15  Dependencia parcial entre las tablas provincias y cantones  Definición formal  Cantones.CodProv=1 and CodCan Between 1 and 20  and Cantones.CodProv =2 and CodCan Between 1 and 15  and Cantones.CodProv =3 and CodCan Between 1 and 8  and Cantones.CodProv =4 and CodCan Between 1 and 10  and Cantones.CodProv =5 and CodCan Between 1 and 11  and Cantones.CodProv =6 and CodCan Between 1 and 11  and Cantones.CodProv =7 and CodCan Between 1 and 6;  Definición informal  Se realiza la dependencia entre cada provincial, con su respectivo número de cantones, la provincia de San José posee 20 cantones, la provincia de Alajuela cuenta con 15 cantones, la provincia de Cartago tiene 8 cantones, Heredia posee 10 cantones, Guanacaste y Puntarenas poseen 11 cantones y Limón cuenta con 6 cantones.  SQL  SELECT \* FROM proyecto2.cantones  where CodProv=1 and CodCan not Between 1 and 20  or CodProv=2 and CodCan not Between 1 and 15  or CodProv=3 and CodCan not Between 1 and 8  or CodProv=4 and CodCan not Between 1 and 10  or CodProv=5 and CodCan not Between 1 and 11  or CodProv=6 and CodCan not Between 1 and 11  or CodProv=7 and CodCan not Between 1 and 6; |

### Screenshot

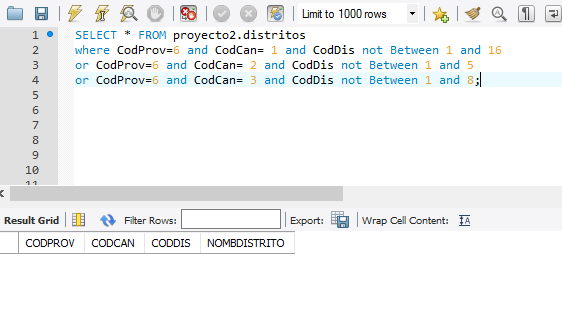


Existe una correcta dependencia parcial entre el cantón y la provincia, es decir, cada provincia, posee un número correcto de cantones, por lo cual no retorna ningún registro.

## Tabla: Provincia-Cantones-Distritos

|  |
| --- |
| Regla #16  Dependencia parcial entre las tablas provincias, cantones y distritos de puntarenas  Definición formal  Distritos.CodProv=6 and CodCan= 1 and CodDis Between 1 and 16  and Distritos.CodProv =6 and CodCan= 2 and CodDis Between 1 and 5  and Distritos.CodProv =6 and CodCan= 3 and CodDis Between 1 and 8;  Definición informal  Se realiza la dependencia entre cada provincial, con su respectivo número de cantones, la provincia de San José posee 20 cantones, la provincia de Alajuela cuenta con 15 cantones, la provincia de Cartago tiene 8 cantones, Heredia posee 10 cantones, Guanacaste y Puntarenas poseen 11 cantones y Limón cuenta con 6 cantones.  SQL  SELECT \* FROM proyecto2.distritos  where CodProv=6 and CodCan= 1 and CodDis not Between 1 and 16  or CodProv=6 and CodCan= 2 and CodDis not Between 1 and 5  or CodProv=6 and CodCan= 3 and CodDis not Between 1 and 8; |

### Screenshot



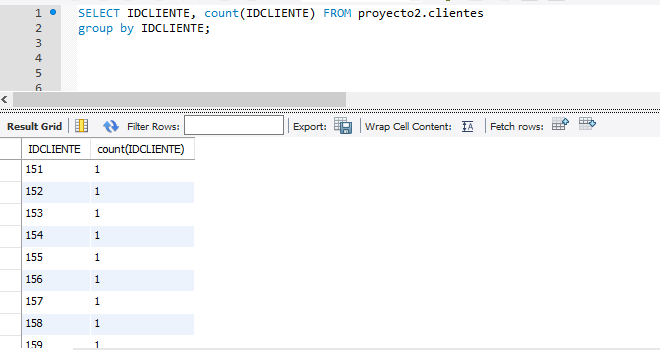
Existe una correcta dependencia parcial entre el cantón, distrito y provincia de Puntarenas, es decir, los cantones de Puntarenas, Esparza y Buenos Aires poseen un correcto número de distritos.

# Restricción de identidad

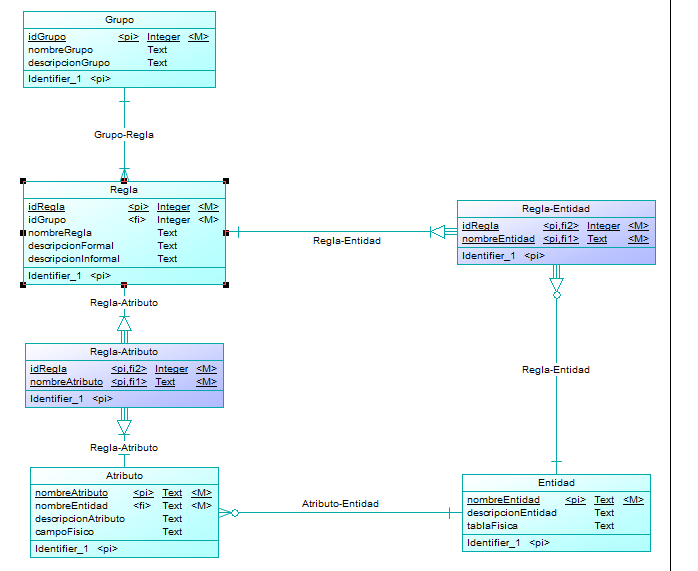
Tabla: Clientes

|  |
| --- |
| Regla #17  Identidad única para cada uno de los clientes  Definición formal  COUNT (Clientes.IDCLIENTE) group by (IDCLIENTE)  Definición informal  Cuenta los registros que poseen el mismo número de ID de Cliente, es caso de que existan duplicados, se mostrará un número superior a dos.  SQL  SELECT IDCLIENTE, count(IDCLIENTE) FROM proyecto2.clientes  group by IDCLIENTE; |

### Screenshot



# Descripción del Diseño de la base de datos del Catalogo

A continuación, se muestra el diseño de la base de datos Catalogo donde se guardará el catálogo de reglas.

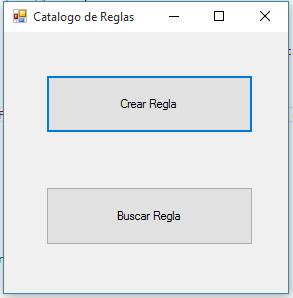
Se decidió crear las tablas Grupo, Atributo, Entidad y Regla, así como las tablas intermedias Regla-Atributo y Regla-Entidad.

La tabla Grupo cuenta con los atributos IdGrupo como llave primaria, nombre y descripción.

La tabla Regla cuenta con los atributos IdRegla como llave primaria, IdGrupo, nombreRegla, descripcionFormal y descripcionInformal.

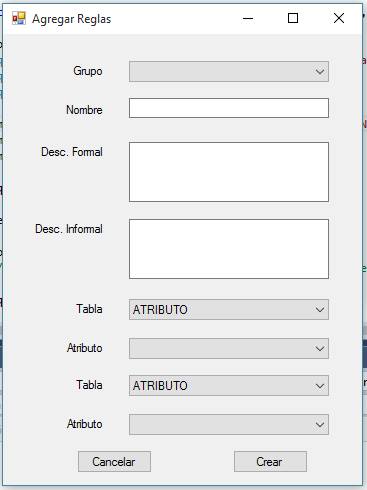
La tabla Entidad cuenta con los atributos nombreEntidad, descripcionEntidad y tablaFisica la cual representa las tablas dentro de la base de datos.

# Diseño y funcionalidades de la aplicación del Catálogo de Reglas.

Al abrir la aplicación se muestra la siguiente pantalla:

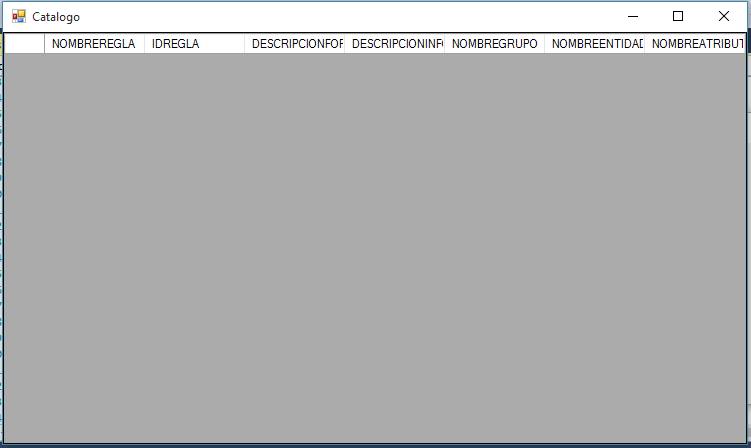
Esta permite seleccionar las funciones de crear nuevas reglas o de buscar reglas existentes.

Al seleccionar la opción de crear nuevas reglas se muestra la siguiente ventana:



Esta permite seleccionar el grupo al que pertenece la regla, el ingreso del nombre, descripción formal e informal y elegir la tabla y los atributos que esta afecta.

La opción de visualizar las reglas muestra la siguiente ventana.



El data grid en esta ventana muestra el catálogo de reglas según la tabla, nombre, idregla, etc.

**Repositorio de código fuente**

El código fuente del proyecto se encuentra en el siguiente enlace:

<https://github.com/VinicioMora/Catalogo-de-Reglas>

**Completitud del proyecto**

El proyecto fue completado al 95%, dado que solo hicieron falta:

* Búsqueda más amigable con el usuario de la aplicación.