

**UNIVERSIDAD DE LIMA**  
**ESCUELA UNIVERSITARIA DE INGENIERÍA**  
**CARRERA DE INGENIERÍA DE SISTEMAS**



***Taller 4: Exploración y Limpieza de Datos***

*Sección 701*

*Curso: SISTEMAS DE INTELIGENCIA EMPRESARIAL*

*Profesor: LUIS ARMANDO RAYGADA VARGAS*

**Integrantes:**

|                                       |                 |
|---------------------------------------|-----------------|
| <b>AGÜERO ESTRELLA AARON FERNANDO</b> | <b>20131537</b> |
| <b>ESPINOZA GRADOS ROSSY</b>          | <b>20131797</b> |
| <b>MANDUJANO NIMA RICARDO JESUS</b>   | <b>20100661</b> |
| <b>PALOMARES HUERTA IVAN ENRIQUE</b>  | <b>20133037</b> |
| <b>ROJAS HINOSTROZA JASON RAFAEL</b>  | <b>20133136</b> |
| <b>UGAZ BURGA CARLOS GABRIEL</b>      | <b>20131349</b> |

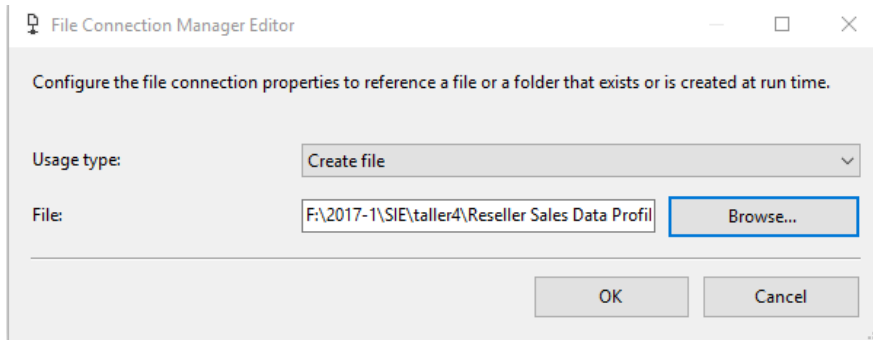
Lima – Perú

Junio 2017

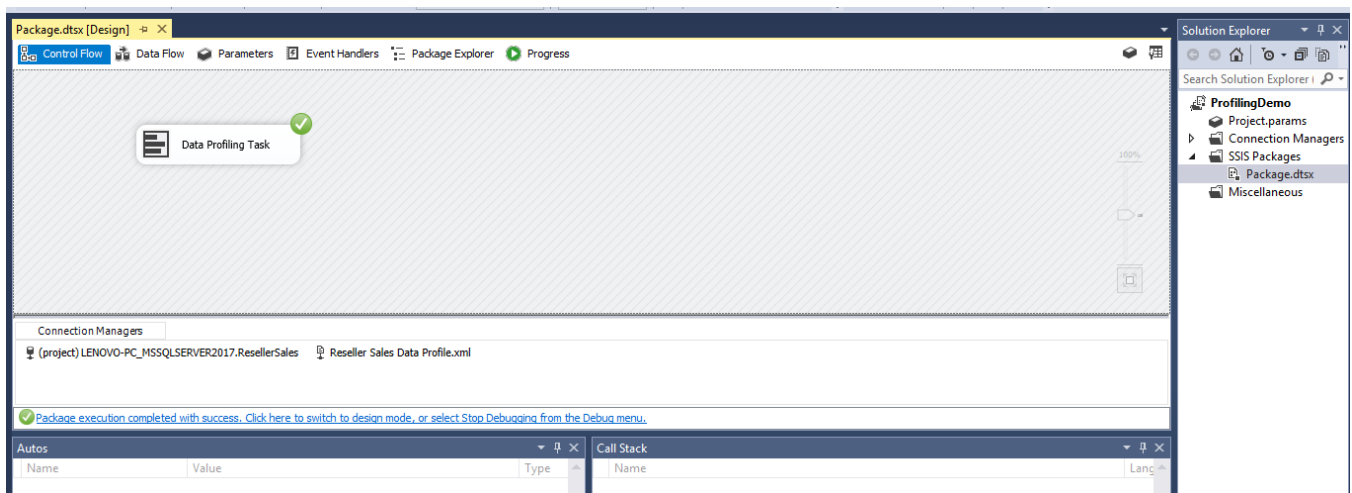
## PARTE SQL SERVER 2014:

Ver los resultados de los perfiles **Value Inclusion Profile Request** y **Column Null Ratio Profile Request**, determinar y escribir que información muestran estos perfiles. En base a la información de la calidad de estos datos ¿Que decisiones tomaría? (se puede ayudar de consultas SQL para tomar más información).

Creación del archivo Reseller Sales Data Profile.xml:



Debugging ejecutado correctamente: Data Profiling Task



| Este equipo > Nuevo vol (F:) > 2017-1 > SIE > taller4 |                      |                     |        |
|---|----------------------|---------------------|--------|
| Nombre  | Fecha de modifica... | Tipo                | Tamaño |
| ProfilingDemo   | 04/06/2017 12:47 ... | Carpeta de archivos |        |
| Reseller Sales Data Profile                           | 04/06/2017 01:04 ... | Archivo XML         | 11 KB  |

### Análisis de resultados:

Data Profile Viewer - F:\2017-1\SIE\tailler4\Reseller Sales Data Profile.xml

Open Refresh

Profiles (Table View)

Column Null Ratio Profiles - [dbo].[Resellers] Encrypted Connection 1000 Rows

Data Sources

- LENOVO-PC\MSSQLSERVER2017
  - Databases
    - ResellerSales
      - Tables
        - [dbo].[Resellers]
          - Column Length Distribution Profiles
          - Column Null Ratio Profiles
          - [dbo].[SalesOrderHeader]
            - Column Statistics Profiles
            - Inclusion Profiles

| Column       | Null Count | Null Percentage |
|--------------|------------|-----------------|
| AddressLine2 | 668        | 95.2924 %       |

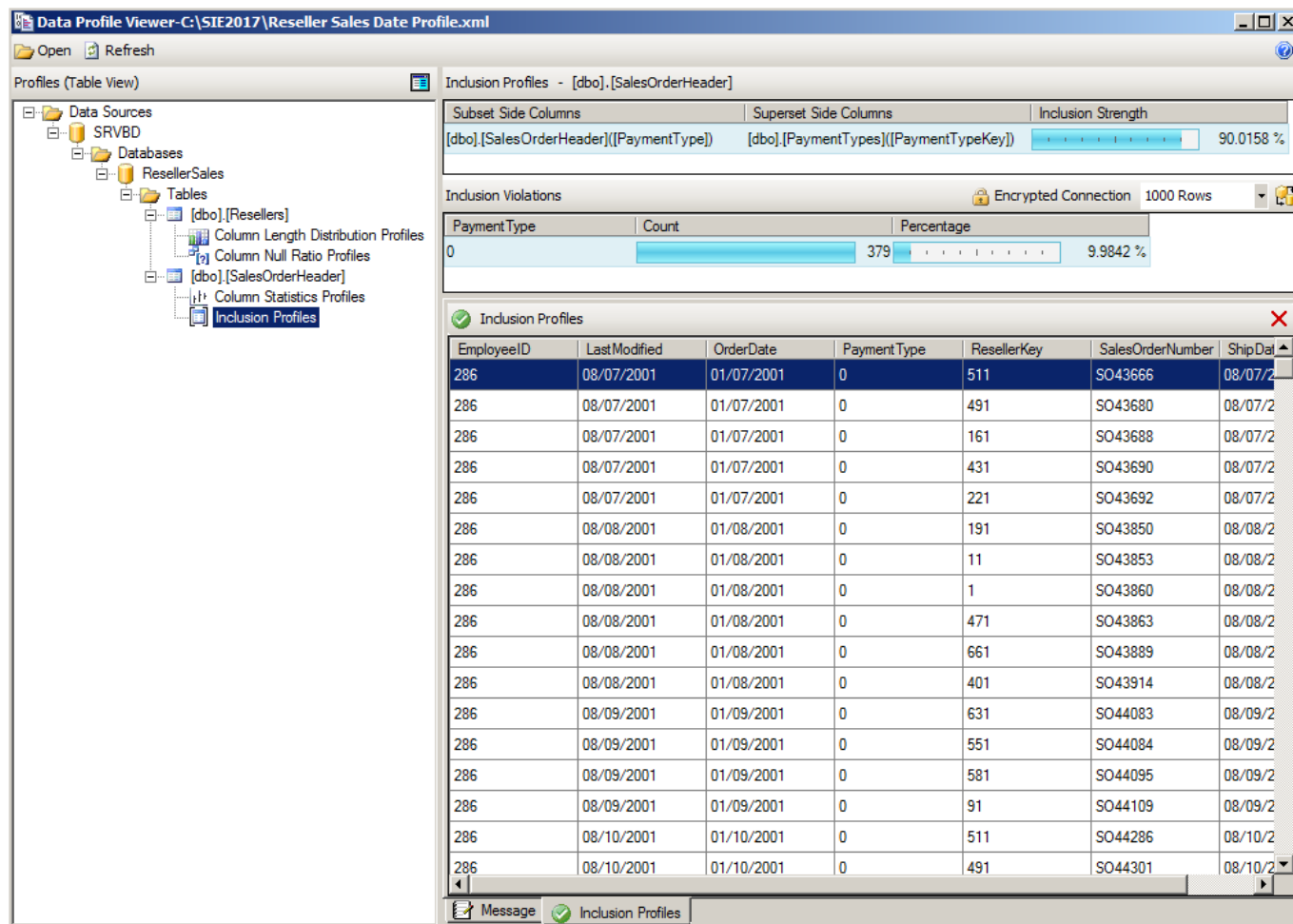
Column Null Ratio Profiles

| AddressLine1         | AddressLine2         | Business Type     | City | Country/RegionCode | Country/RegionNam | NumberEmployees     | Phone   | PostalCode | Reseller |
|----------------------|----------------------|-------------------|------|--------------------|-------------------|---------------------|---------|------------|----------|
| 2251 Eliot Avenue    | Value Added Res...   | Seattle           | US   | United States      | 2                 | 245-555-0173        | 98104   | 1          |          |
| 3207 S Grady Way     | Specialty Bike Sh... | Renton            | US   | United States      | 10                | 170-555-0127        | 98055   | 2          |          |
| 12345 Sterling A...  | Warehouse            | Irvine            | US   | United States      | 40                | 279-555-0130        | 75061   | 3          |          |
| 482505 Warm Sp...    | Specialty Bike Sh... | Fremont           | US   | United States      | 13                | 828-555-0186        | 94536   | 5          |          |
| 39933 Mission O...   | Warehouse            | Camarillo         | US   | United States      | 43                | 244-555-0112        | 93010   | 6          |          |
| 5420 West 2250...    | Value Added Res...   | Salt Lake City    | US   | United States      | 8                 | 192-555-0173        | 84101   | 7          |          |
| 79945 Corporate ...  | Specialty Bike Sh... | Miami             | US   | United States      | 16                | 872-555-0171        | 33127   | 8          |          |
| 3333 Micro Drive     | Warehouse            | Millington        | US   | United States      | 46                | 488-555-0130        | 38054   | 9          |          |
| 6388 Lake City ...   | Value Added Res...   | Burnaby           | CA   | Canada             | 11                | 150-555-0127        | V5A 3A6 | 10         |          |
| 52560 Free Street    | Specialty Bike Sh... | Toronto           | CA   | Canada             | 19                | 926-555-0159        | M4B 1V7 | 11         |          |
| 22580 Free Street    | Warehouse            | Toronto           | CA   | Canada             | 49                | 112-555-0191        | M4B 1V7 | 12         |          |
| 7033, rue de Lon...  | Value Added Res...   | Les Ulis          | FR   | France             | 14                | 1 (11) 500 555-0... | 91940   | 13         |          |
| Karl Liebknecht s... | Specialty Bike Sh... | Frankfurt am Main | DE   | Germany            | 22                | 1 (11) 500 555-0... | 60075   | 14         |          |
| 25 Epping Road       | Warehouse            | Lavender Bay      | AU   | Australia          | 52                | 1 (11) 500 555-0... | 2060    | 15         |          |
| 93-2501, Blackfri... | Value Added Res...   | London            | GB   | United Kingdom     | 17                | 1 (11) 500 555-0... | SE1 8HL | 16         |          |
| 9920 Picketts Lin... | Specialty Bike Sh... | Newport News      | US   | United States      | 25                | 497-555-0147        | 23607   | 17         |          |

Message Column Null Ratio Profiles

En esta parte se visualiza que el 95.29% de la información de la columna AddressLine2 en la tabla Resellers es nula.

Para solucionar este inconveniente, se haría una transformación de columna para indicar “No tiene” en caso de que el registro tenga valor nulo en el campo AddressLine2. Por ello, se aplicaría Derived Column con la función ReplaceNull para el campo AddressLine2.

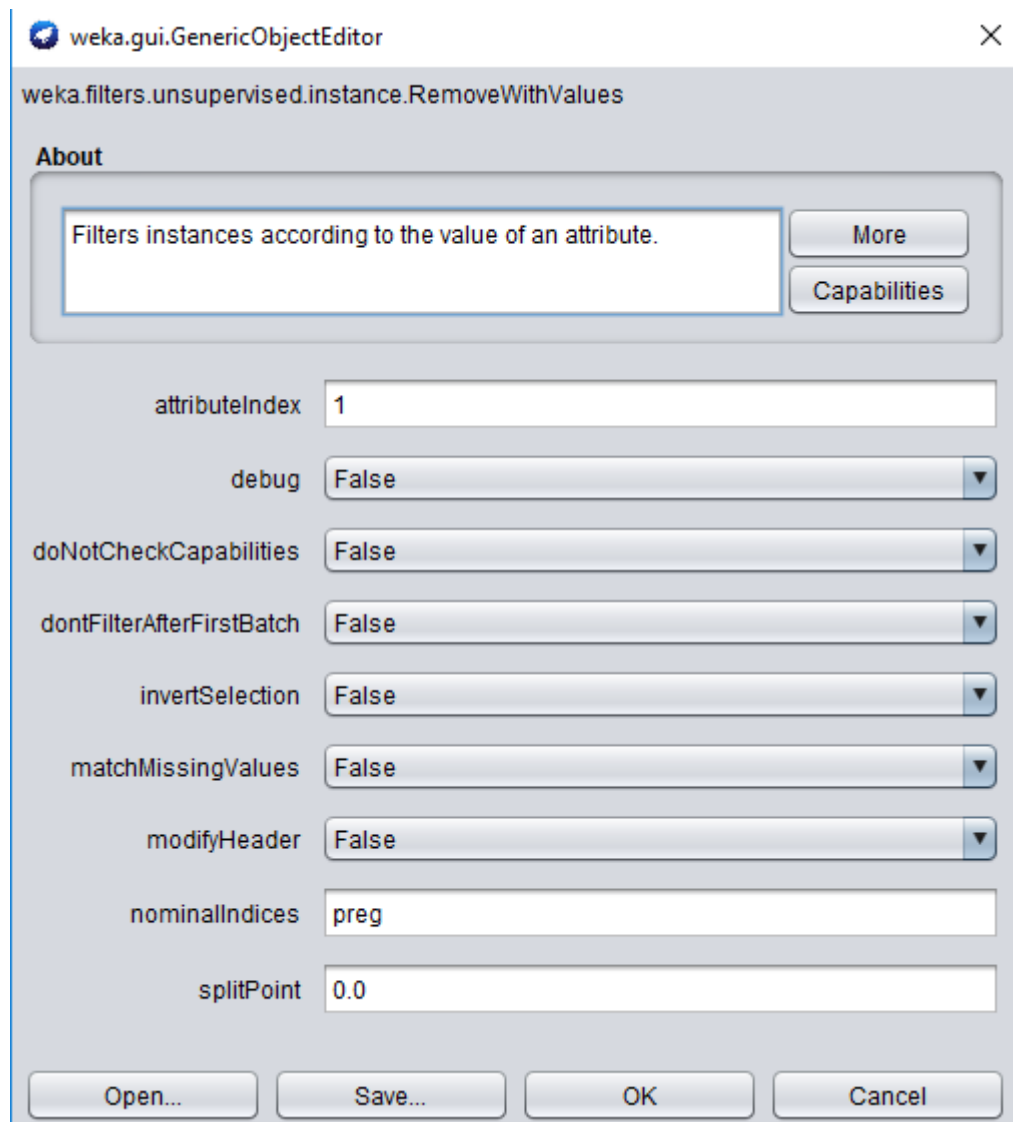


En esta imagen se visualiza que el 90.01% de los datos de la columna PaymentType de la tabla SalesOrderHeader hace referencia a la columna PaymentTypeKey de la tabla PaymentTypes. Sin embargo, el 9.98% restante de la columna PaymentType no hace referencia a la columna de la tabla PaymentTypes. Para ello se realizaría una función Lookup con una consulta de inner joins que obtenga solamente los registros que hagan referencia a la tabla de PaymentTypes.

## PARTE WEKA – DIABETES:

Primero, se separó en 6 rangos los valores del atributo edad.

1. Se eliminó los registros que contenían el valor -1 como número de embarazos.



Open file...

Open URL...

Open DB...

Generate...

Undo

Edit...

Save...

## Filter

Choose

RemoveWithValues -S 0.0 -C 1 -L preg

Apply

## Current relation

Relation: pima\_diabetes-weka.filte...  
Instances: 763Attributes: 9  
Sum of weights: 763

## Attributes

All

None

Invert

Pattern

| No. | Name                                     |
|-----|--|
| 1   | <input checked="" type="checkbox"/> preg |
| 2   | <input type="checkbox"/> plas            |
| 3   | <input type="checkbox"/> pres            |
| 4   | <input type="checkbox"/> skin            |
| 5   | <input type="checkbox"/> insu            |
| 6   | <input type="checkbox"/> mass            |
| 7   | <input type="checkbox"/> pedi            |

Remove

## Selected attribute

Name: preg  
Missing: 0 (0%)

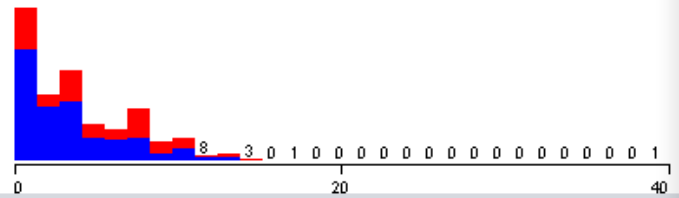
Distinct: 18

Type: Numeric  
Unique: 3 (0%)

| Statistic | Value |
|-----------|-------|
| Minimum   | 0     |
| Maximum   | 40    |
| Mean      | 3.907 |
| StdDev    | 3.599 |

Class: class (Nom)

Visualize All



## Status

OK

Log

x 0

2. Se reemplazó los valores vacíos por la moda y la media pues el campo plasma es un tipo numérico. Además, al utilizar estas medidas de tendencia central, la alteración de la variabilidad es mínima.

Weka Exp

weka.gui.GenericObjectEditor

Preprocess

Open file

Filter

Choose

Current relation

Relation Instances

Attributes

All

No.

1

2

3

4

5

6

7

skin

insu

mass

pedi

Remove

debug False

doNotCheckCapabilities False

ignoreClass False

Open...

Save...

OK

Cancel

Visualize All

About

Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.

More

Capabilities

Save...

Apply

Visualize All

Status

OK

Log

x 0

| Value | Frequency |
|-------|-----------|
| 5     | 5         |
| 0     | 0         |
| 1     | 1         |
| 5     | 5         |
| 16    | 16        |
| 61    | 61        |
| 108   | 108       |
| 135   | 135       |
| 128   | 128       |
| 89    | 89        |
| 73    | 73        |
| 44    | 44        |
| 41    | 41        |
| 29    | 29        |
| 28    | 28        |

8. Al tener los 6 rangos de edad, se reemplazaron los rótulos que indicaban valores infinitos. Los cambios fueron los siguientes: '(0-31]' en vez de '(-inf-31]' y '(71-130]' en reemplazo de '(71-inf]'.

