

**Laboratorio N°4**

***Exploración y Limpieza de Datos***

Profesores Responsables:

**Ing. Edgardo Lazo**

**Ing. Luis Raygada**

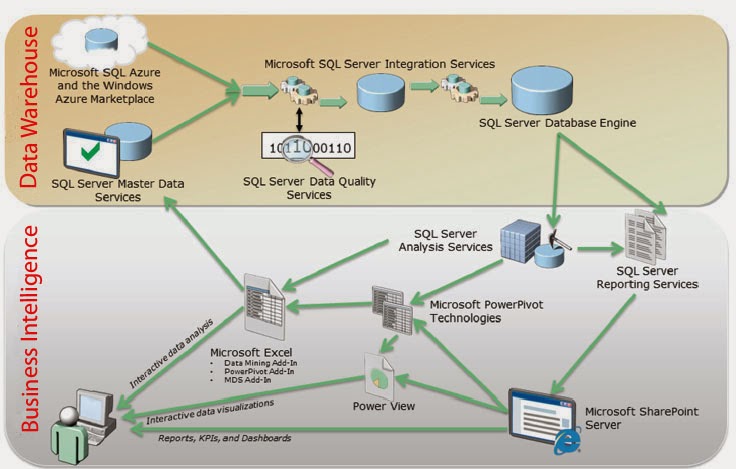
**SISTEMAS DE INTELIGENCIA EMPRESARIAL**

**Objetivos:**

***Este material es de apoyo académico y se hace para uso exclusivo de los alumnos de la Universidad de Lima y en concordancia con lo dispuesto por la legislación sobre los derechos de autor: Decreto Legislativo 822***

**OBJETIVOS DEL LABORATORIO:**

* Consolidar los conocimientos adquiridos en Inteligencia de Negocios y plasmarlos en MS Sql Server.
* Reconocer el entorno de trabajo de SQL.
* Identificar las distintas transacciones y componentes.
* Reconocer el entorno de trabajo de Weka



**LABORATORIO #1**

* **En la presente guía se detallan los siguientes puntos:**
  + Marco teórico general de los puntos a desarrollar.
  + Capturas de pantallas para reconocimiento del entorno de trabajo.
  + Desarrollo y configuración.

**Fuente:**

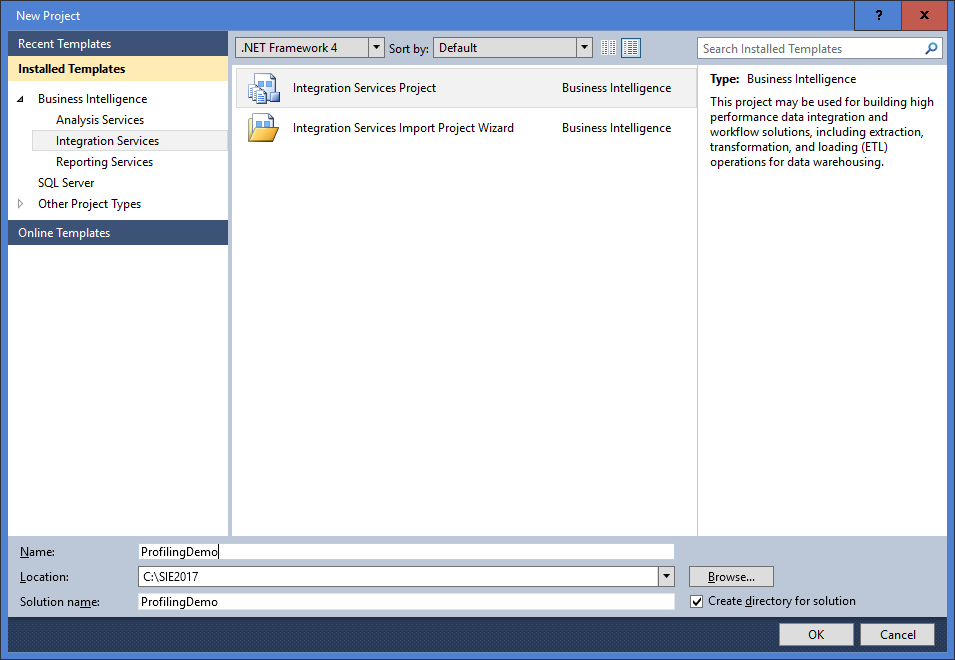
* + SQLServer 2014
  + Material de clase.
  + Weka

***Los profesores de la asignatura.***

***Tarea 1: Crear un Perfil de Datos***

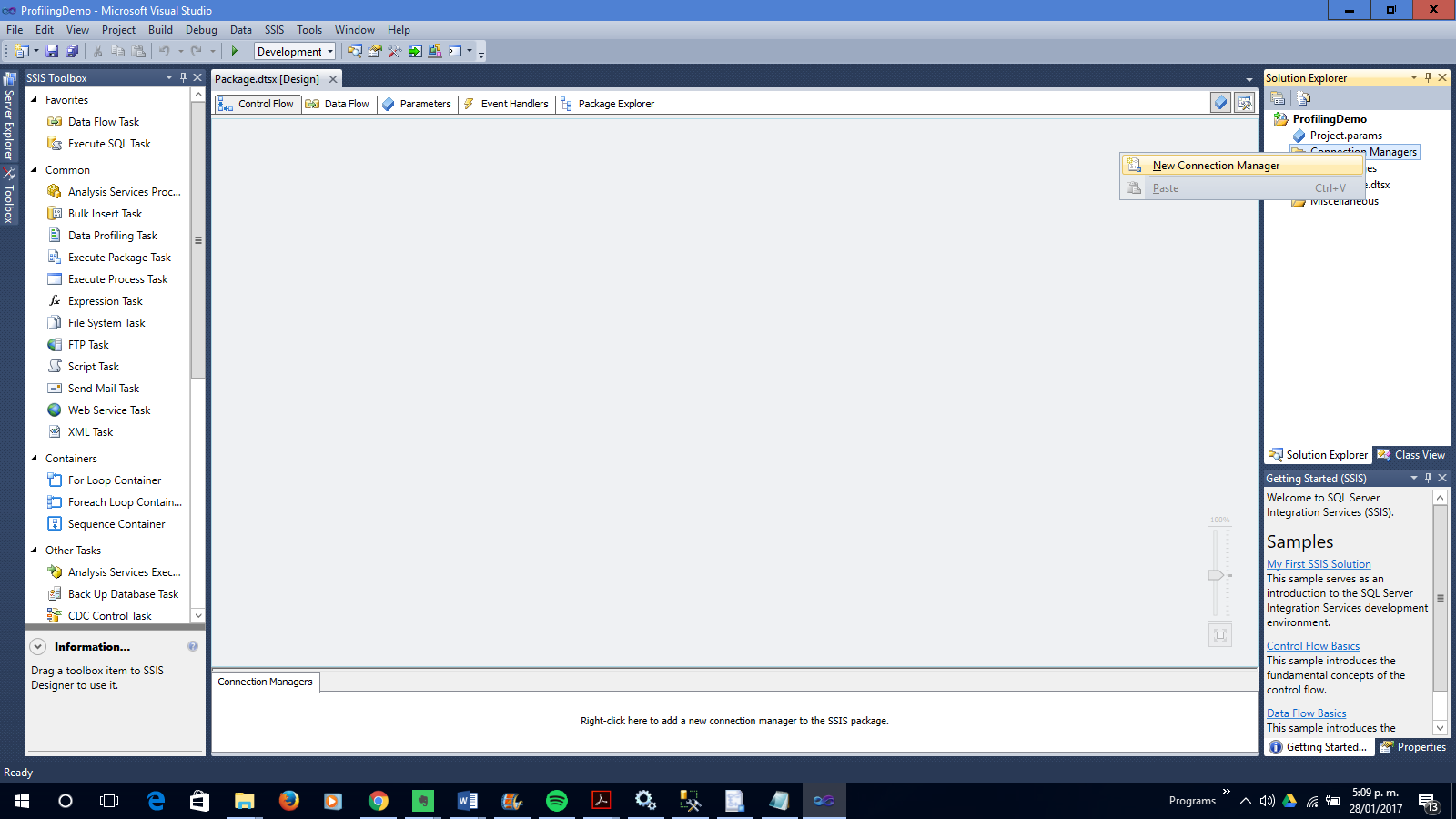
Hacer un restore de la base de datos Reseller Sales que será proporcionada por el profesor.

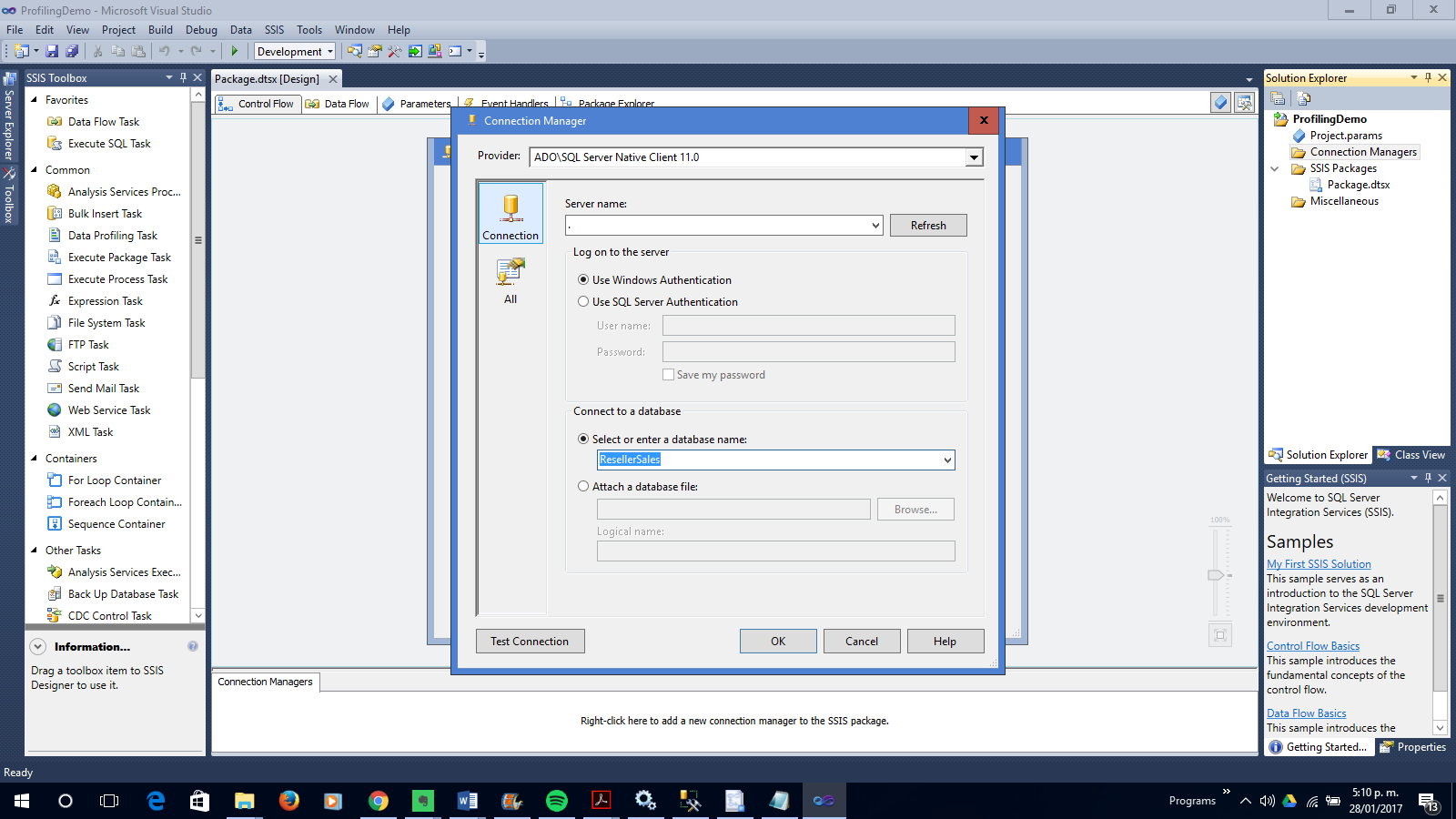
Ingrese a SQL Server Data Tools y cree un nuevo proyecto (File – New – Project) de Integration Services denominado ProfilingDemo en la carpeta C: \SIE2017.



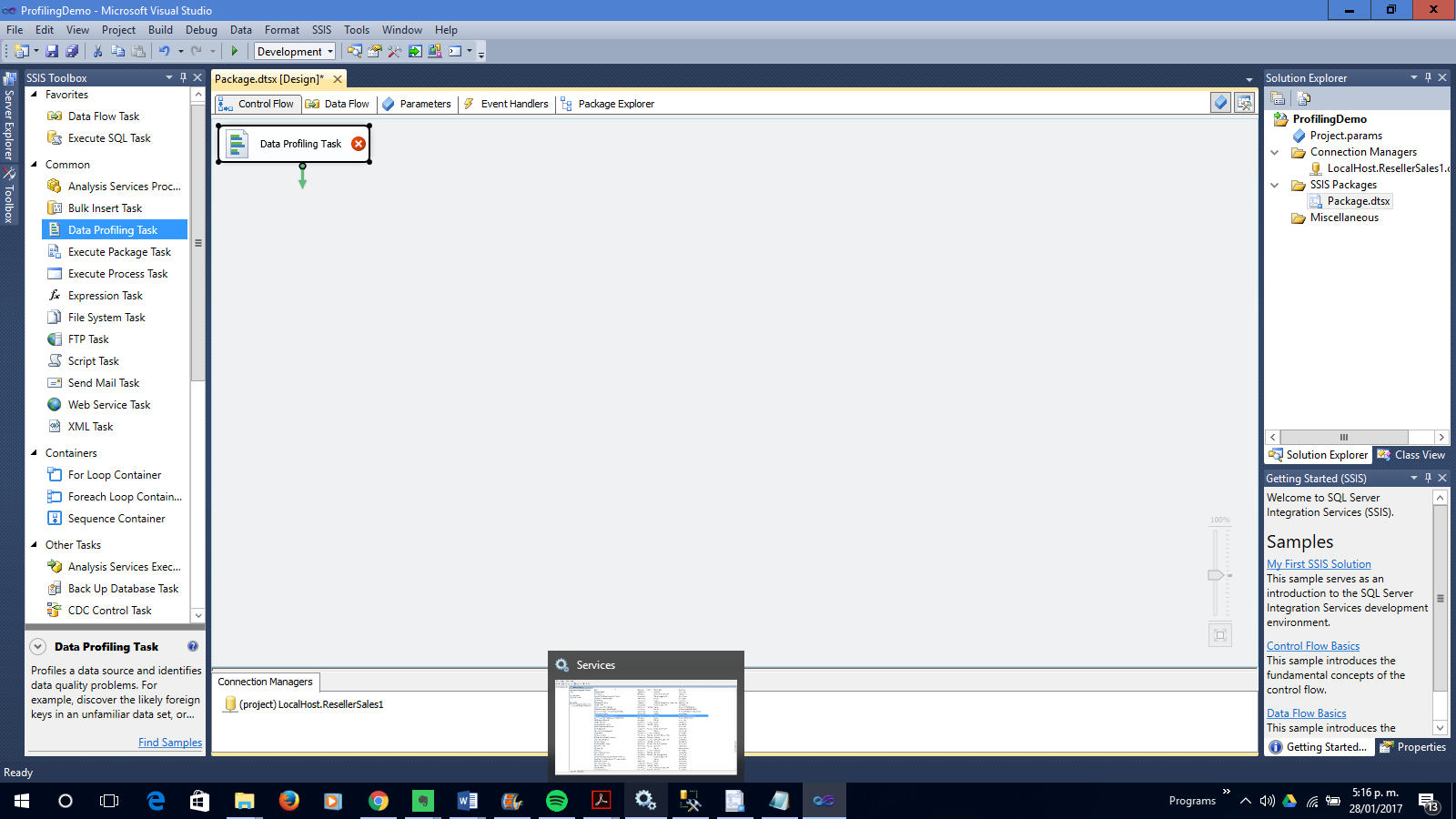
3. En el panel Solution Explorer, cree un nuevo gestor de conexiones ADO.NET con la configuración siguiente:

* Nombre del servidor: localhost
* Inicie sesión en el servidor: Utilizar autenticación de Windows
* Base de datos: ResellerSales





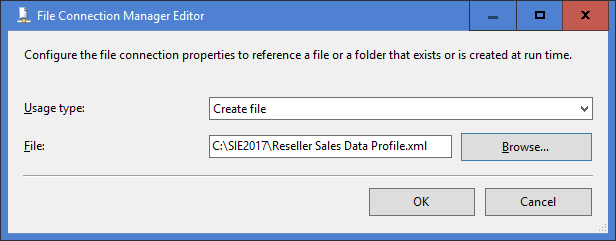
4. En el panel SSIS Toolbox, en la sección **Common**, haga doble clic en **Data Profiling Task** para agregar la tarea al área Control Flow.



1. Haga doble clic en el icono **Data Profiling Task**. En **Data Profiling Task Editor** en la ficha **General**, en la propiedad **Destination** seleccione **<New File connection…>**.

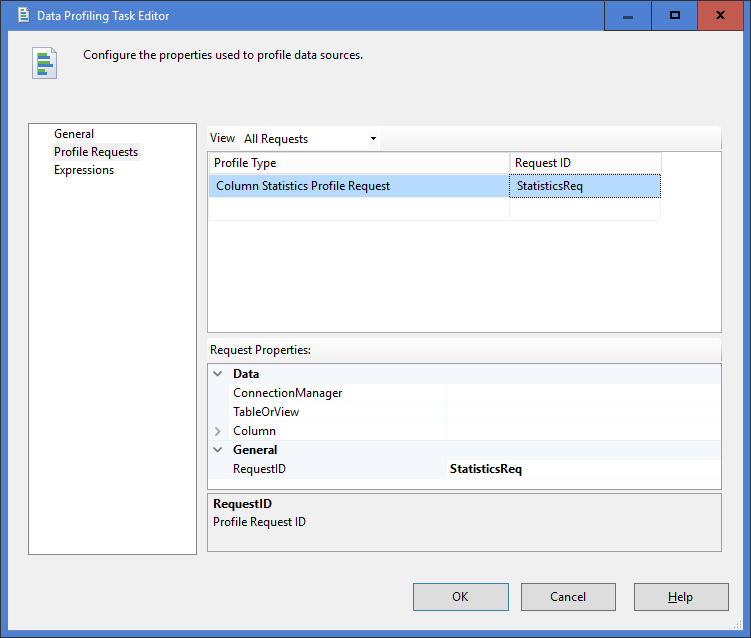
En la lista desplegable **Usage type**, haga clic en Create File.

En el cuadro File, escriba c:\SIE2017\Reseller Sales Data Profile.xml y, a continuación, haga clic en Aceptar.



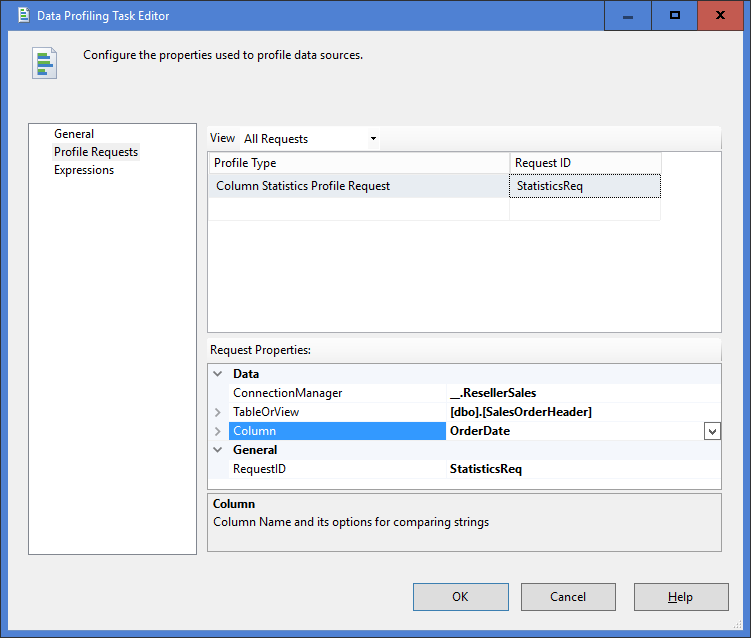
9. En la ficha General, establezca OverwriteDestination en True.

10. Selecciona la ficha **Profile Requests**, en la lista desplegable **Profile Type,** selecciona **Column Statistics Profile Request** y, a continuación, haga clic **RequestID.**



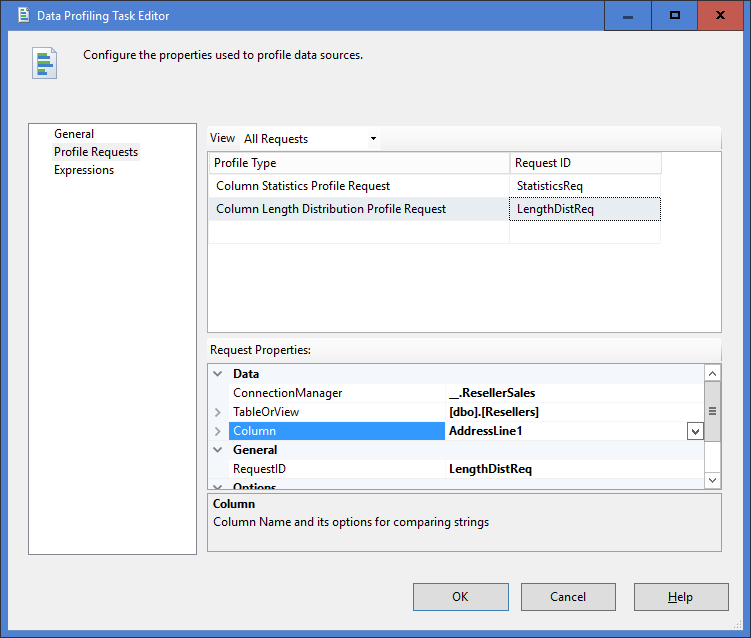
En el panel Request Properties, establezca los siguientes valores de propiedad. No haga clic en Aceptar cuando haya terminado:

* + **ConnectionManager**: localhost.ResellerSales
  + **TableOrView**: [dbo].[SalesOrderHeader]
  + **Column**: OrderDate



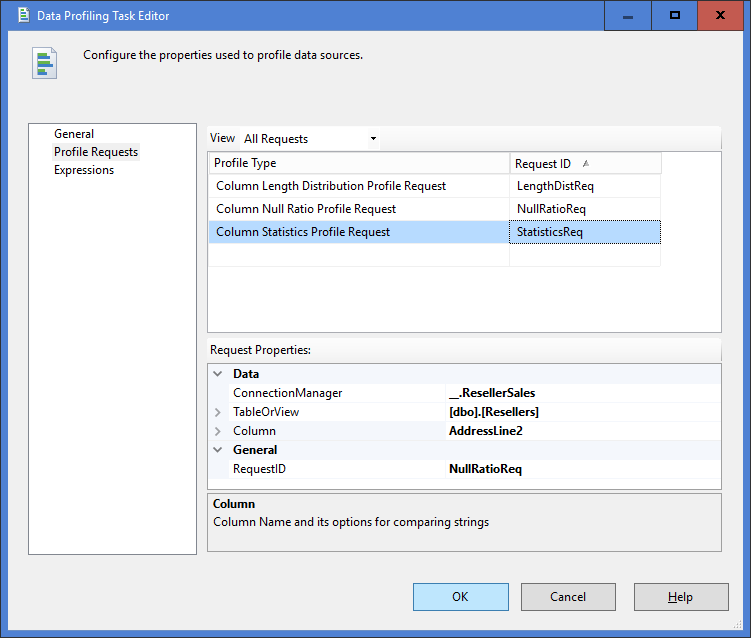
11. Haga click debajo de la fila **Column Statistics Profile Request**, agregue el perfil **Column Length Distribution Profile Request** con las siguientes opciones:

* + **ConnectionManager**: localhost.ResellerSales
  + **TableOrView**: [dbo].[Resellers]
  + **Column**: AddressLine1



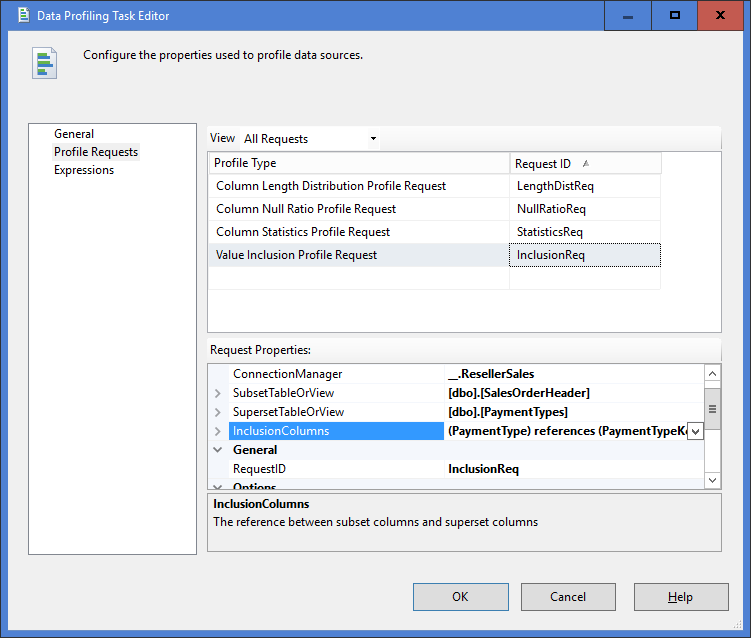
12. Agregue el tipo de perfil **Column Null Ratio Profile Request** con las siguientes configuraciones:

* + **ConnectionManager**: localhost.ResellerSales
  + **TableOrView**: [dbo].[Resellers]
  + **Column**: AddressLine2



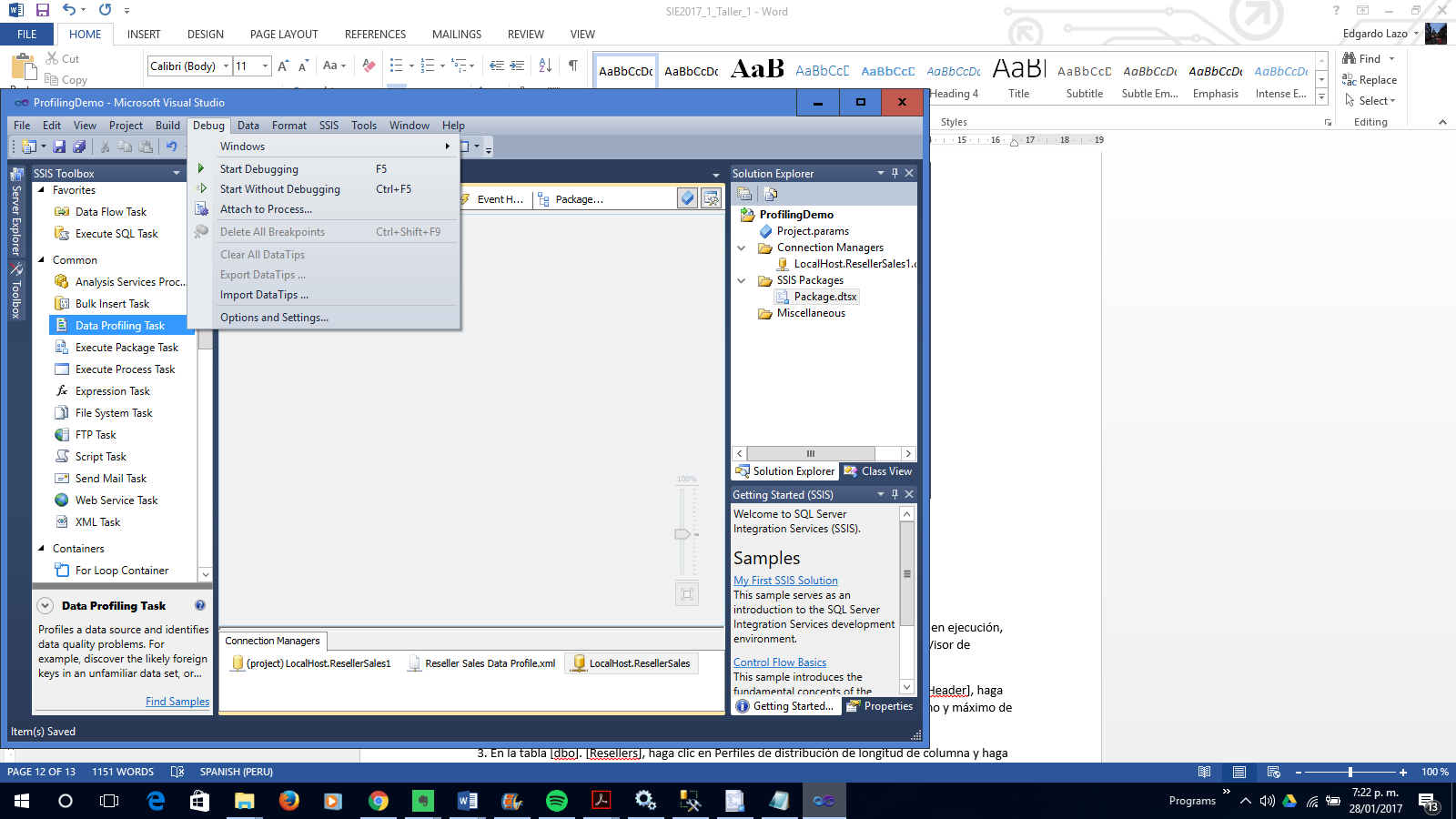
13. Agregue el tipo de perfil **Value Inclusion Profile Request** con las siguientes configuraciones:

* + **ConnectionManager**: localhost.ResellerSales
  + **SubsetTableOrView**: [dbo].[SalesOrderHeader]
  + **SupersetTableOrView**: [dbo].[PaymentTypes]
  + **InclusionColumns**:
    - *Subset side Columns*: PaymentType
    - *Superset side Columns*: PaymentTypeKey
  + **InclusionThresholdSetting**: None
  + **SupersetColumnsKeyThresholdSetting**: None
  + **MaxNumberOfViolations**: 100



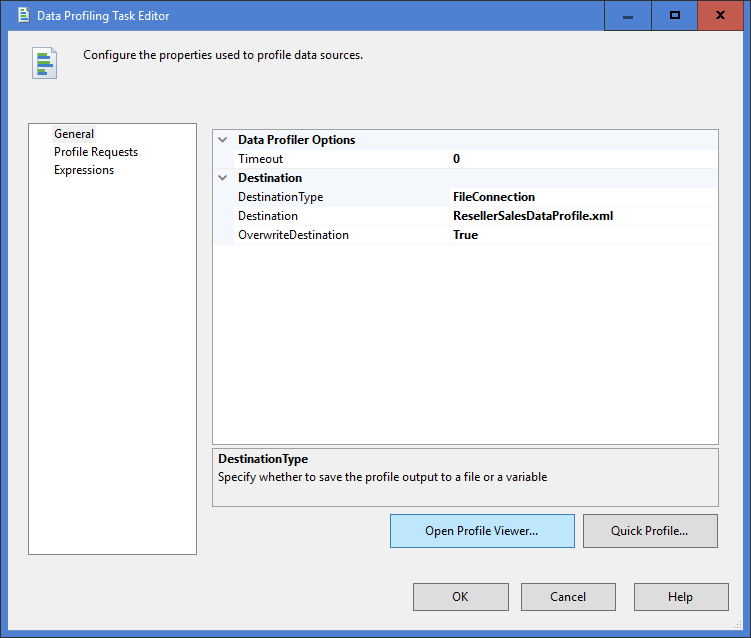
14. De click en Aceptar/Ok.

16. En el menú seleccione **Debug**, luego **Start Debugging**.



***Tarea 3: Ver un informe de perfil de datos***

1. Cuando se haya completado la tarea Data Profiling task, de doble click a **Data Profiling** y, a continuación, haga clic en **Open Profile Viewer**.

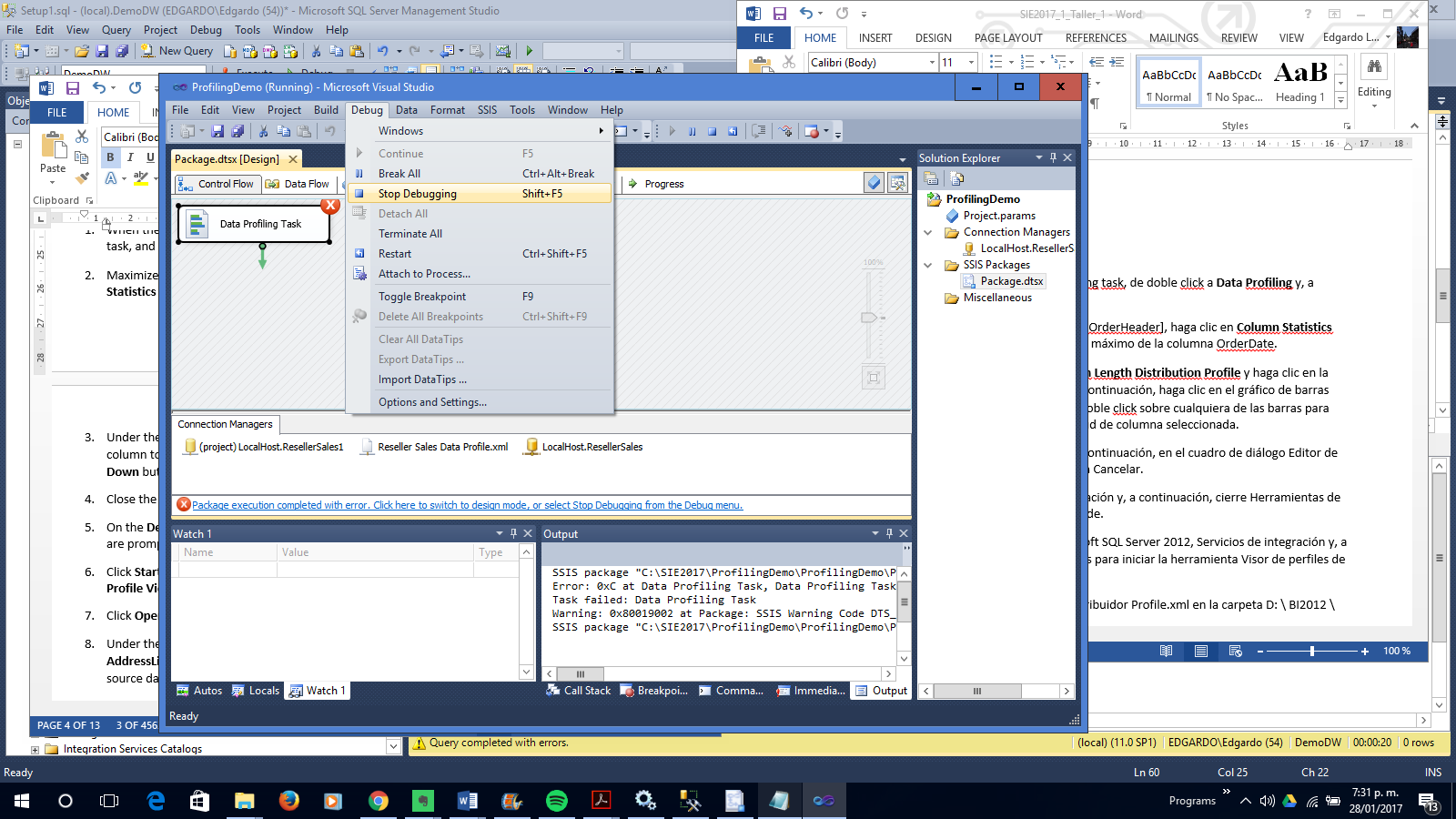


2. Maximice la ventana y, bajo la tabla [dbo]. [SalesOrderHeader], haga clic en **Column Statistics Profile**. A continuación, revise los valores mínimo y máximo de la columna OrderDate.

3. En la tabla [dbo]. [Resellers], haga clic en **Column Length Distribution Profiles** y haga clic en la columna AddressLine1 para ver las estadísticas. A continuación, haga clic en el gráfico de barras para cualquiera de las longitudes de columna, de doble click sobre cualquiera de las barras para ver los datos de origen que coinciden con la longitud de columna seleccionada. **SI COINCIDEN**

4. Cierre la ventana Visor de perfiles de datos y la ventana Data Profiling task Editor.

5. En el menú Debug, haga clic en Stop Debugging y, a continuación, cierre Herramientas de datos de SQL Server, guardar los cambios si se le pide.



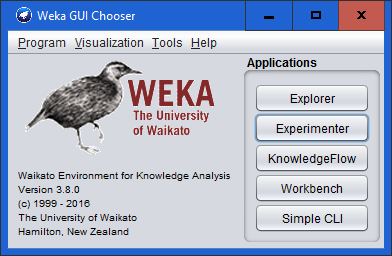
6. Haga clic en Inicio, Todos los programas, Microsoft SQL Server 2012, y a continuación, haga clic en Data Profile Viewer para iniciar la herramienta Visor de perfiles de datos.

7. Haga clic en Open y abra el archivo c:\SIE2017\Reseller Sales Data Profile.xml.

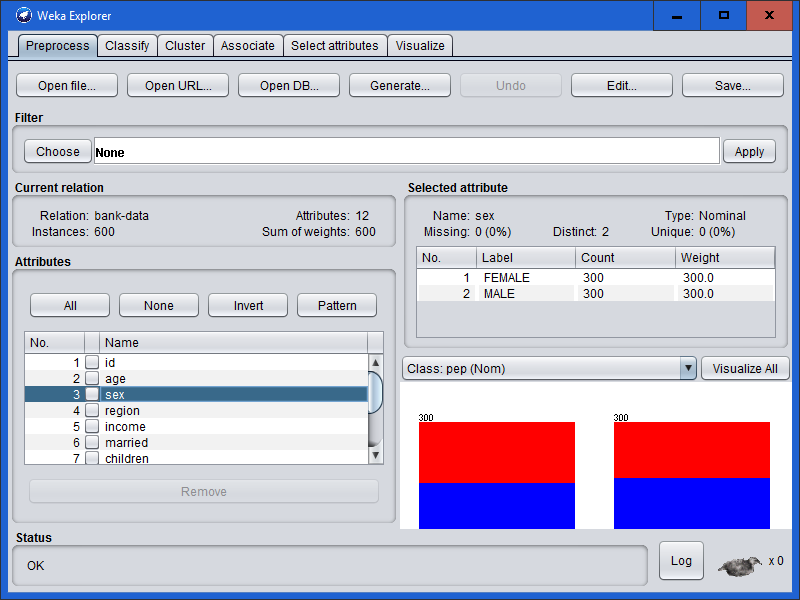
8. Ver los resultados de los perfiles **Value Inclusion Profile Request y Column Null Ratio Profile Request,** determinar y escribir que información muestran estos perfiles. En base a la información de la calidad de estos datos ¿Que decisiones tomaría? (se puede ayudar de consultas SQL para tomar más información). Escriba sus respuestas en un documento Word con nombre NombreApellido.doc y envíalo por mensaje de blackboard al profesor junto con el archivo Reseller Sales Data Profile.xml.

***Tarea 1: Limpieza de datos usando Weka***

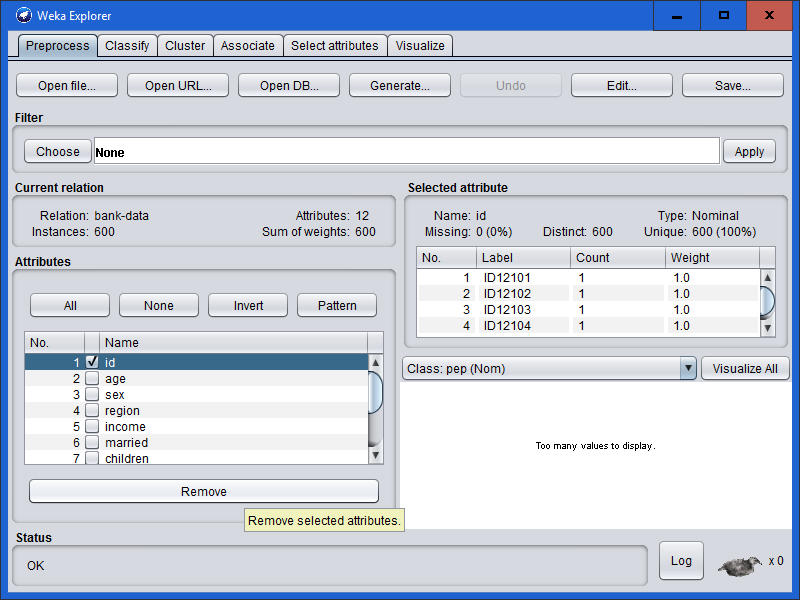
1. Ingrese al aplicativo Weka y seleccione “Explorer”



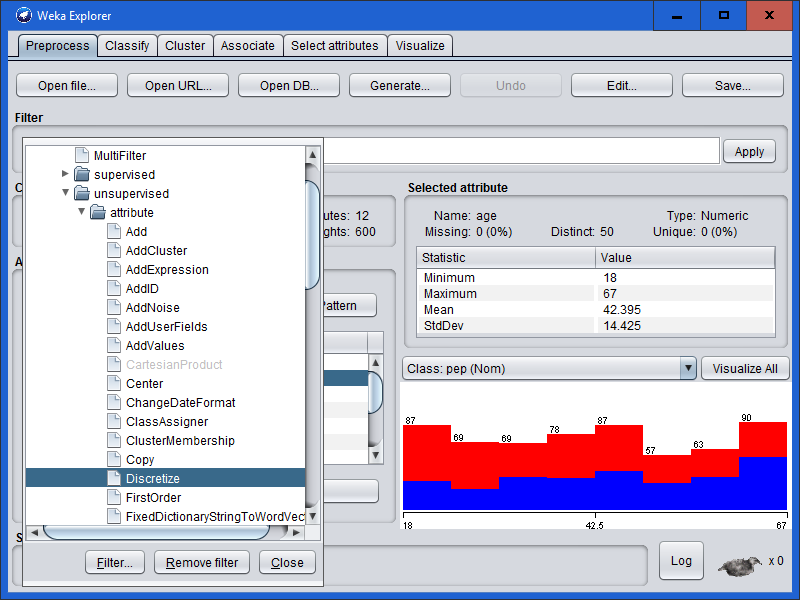
1. Selecciones Openfile y abra el archivo bank-data.arff que será entregado por el profesor. Este dataset muestra los diferentes atributos que se evalúan en un banco para determinar si se da crédito a una persona. Seleccione cada uno de los atributos y vea sus características.



1. Dado que el atributo “id” no aporta información, elimínelo. Seleccione el atributo y de click en Remove.

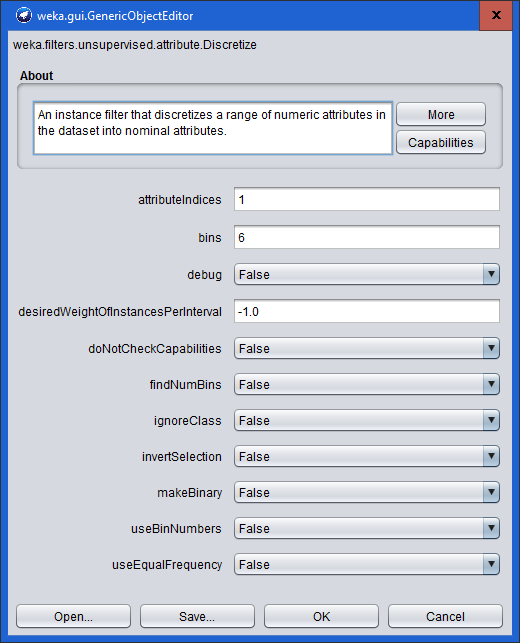


1. Ahora vamos a discretizar el atributo age, es decir lo vamos a convertirlo a rangos de edad. Seleccione el atributo age y debajo de Filter, de click en Choose para seleccionar un algoritmo de filtrado. Seleccione WEKA.filters.unsupervised.attribute.discretize

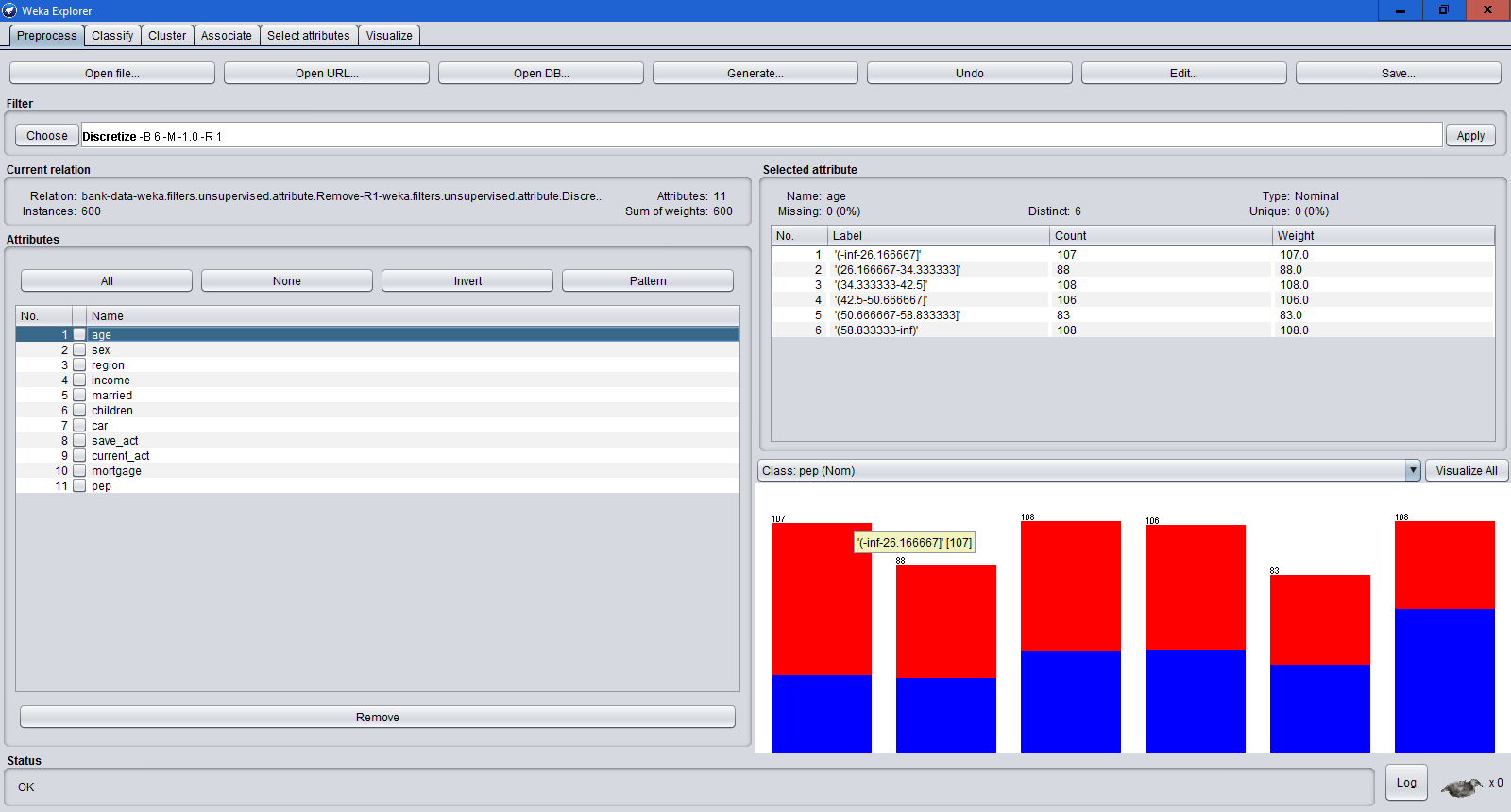


1. Da click en “Discretize ….” y podras ver los parámetros de configuración. Coloca los siguiente parámetros:
   * Attribute índices = 1 (se refiere al atributo age)
   * Bins = 6 (para dividir el atributo age en 6 intervalos)

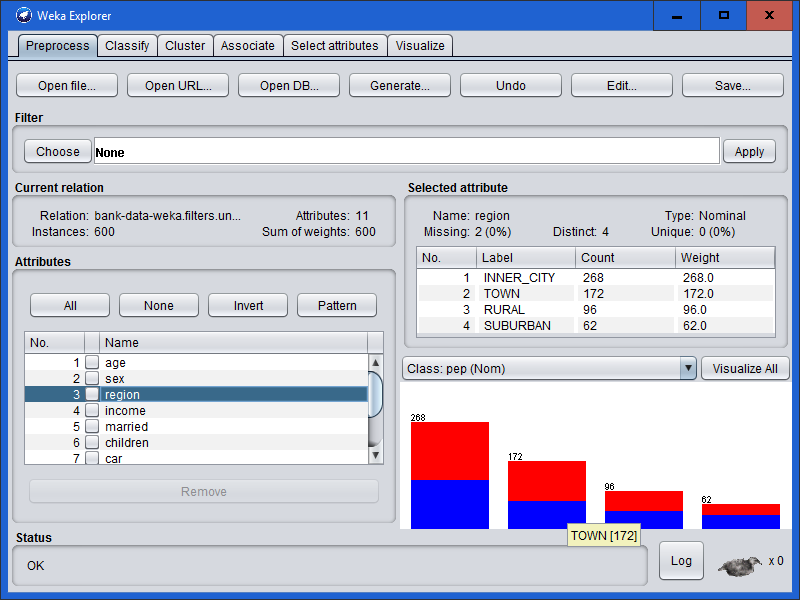
De click en Ok



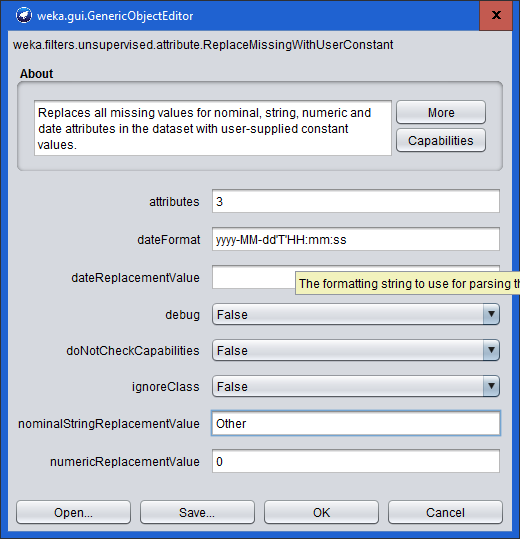
1. Regresará a la pantalla principal, donde se mostrarán los parámetros configurados. De click en Apply.



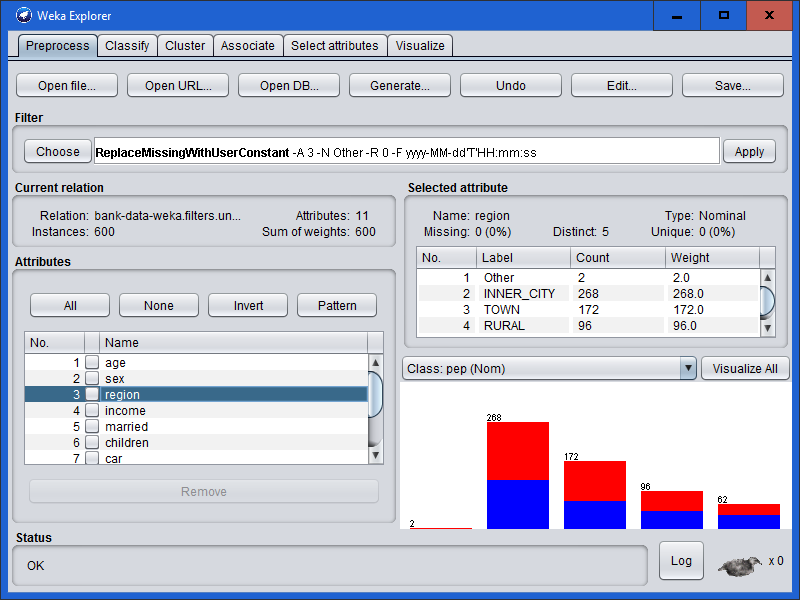
1. Seleccione el atributo Region. En la pantalla Selected atribute se puede ver que este atributo tiene 2 datos perdidos.



1. En el menú filters seleccione el Filtro WEKA.filters.unsupervised.attribute.replacemissingvaluewithuserconstant. Configure como sigue, para que se reemplacen los valores perdidos por “Other”:



1. Aplique el filtro seleccionado.



**Abra el archivo diabetes.arff proceda y agrupe a las personas por 6 rangos de edad.** Luego revise los datos campo por campo y elimine/cambie datos irregulares/anómalos. Si va a reemplazar un valor nulo, evalué cual es el mejor valor a usar para el reemplazo (mínimo, máximo, promedio, cero, etc). Para eliminar los valores anómalos puede usar la función no supervisada “removewithvalues”.

Describa los cambios realizados en un documento Word con nombre NombreApellido.doc y envíalo por mensaje de blackboard al profesor junto con el archivo diabetes.arff ya procesado.

La descripción de los atributos del dataset es:

% 1. preg: Number of times pregnant

% 2. plas: Plasma glucose concentration a 2 hours in an oral glucose tolerance test

5 valores missing:

WEKA.filters.unsupervised.attribute.replacemissingvaluewithuserconstant. Escribir 0.

% 3. pres: Diastolic blood pressure (mm Hg)

% 4. skin: Triceps skin fold thickness (mm)

% 5. insu: 2-Hour serum insulin (mu U/ml)

% 6. mass: Body mass index (weight in kg/(height in m)^2)

% 7. pedi: Diabetes pedigree function

% 8. age: Age (years)

% 9. class: Class variable (0 or 1)