**MÓDULO 5 – PRÁCTICA FINAL**

**BUSINESS INTELIGENCE**

1. **Análisis de fuentes:**
   * + **Descripción global de las fuentes.**
     + **Descripción en detalle de cada campo.**
     + **Tipo de campo, naturaleza, cardinalidad aproximada.**

Para el presente proyecto por desarrollar, hemos recibido un archivo con la información pertinente a una evaluación realizada en España por el departamento antifraude de la compañía de Mystery Shopping que desea hacer un seguimiento y analizar la información relativa a las encuestas que realizan en sus distintos centros operativos.

Partiendo desde la fuente de datos que nos fue entregada (archivo .csv), realizaremos un análisis exploratorio de los datos, determinamos la calidad de éstos, planteamos los hallazgos para discutir la mejor forma de realizar la posible limpieza de datos y tener una fuente confiable para que el análisis a realizar provea de los resultados esperados.

El archivo consiste en 32797 registros, y 12 columnas (o atributos):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Atributo** | **Valores** | **Nulos** | **Únicos** | **Min** | **Max** | **Tipo** |
| OFICINA | 32797 |  | 13 |  |  | Caracter |
| PROVINCIA | 32797 |  | 53 |  |  | Caracter |
| POBLACION | 32716 | 81 | 1654 |  |  | Caracter |
| CP | 32797 |  | 1797 |  |  | Caracter |
| COD\_LOC | 32797 |  | 5645 |  |  | Caracter |
| NOMBRE\_LOC | 32797 |  | 6555 |  |  | Caracter |
| ID\_EVALUACION | 32797 |  | 32797 | 1,938,117 | 2,016,756 | Numérico |
| COD\_PROY | 32797 |  | 222 |  |  | Caracter |
| COD\_AUDITOR | 32797 |  | 1144 |  |  | Caracter |
| TITULO\_CUESTIONARIO | 32797 |  | 439 |  |  | Caracter |
| RESULTADO | 32797 |  | 405 | 0.0 | 1.0 | Numérico |
| FECHA\_DE\_EJECUCION | 32796 | 1 | 188 | 1995-05-03 | 2014-12-12 | Fecha |

Después de una revisión de los datos, la interpretación de éstos es la siguiente:

* COD\_LOC:

Código de referencia ubicación de localización donde se realiza la encuesta

* NOMBRE\_LOC:

Descripción de la referencia de ubicación de la localización donde se realiza la encuesta

* CP:

Código Postal de la ubicación donde se realiza la encuesta

* POBLACION:

Población donde se realiza la encuesta

* OFICINA:

Oficina o sucursal que realiza la encuesta

* PROVINCIA:

Donde se realiza la encuesta

* COD\_PROY:

Identificador del proceso de evaluación

* ID\_EVALUACION:

Identificador único de cada encuesta, por lo tanto, la llave primaria para el control de información

* FECHA\_DE\_EJECUCION:

Fecha de realización de la encuesta

* COD\_AUDITOR:

Identificador del auditor que realiza la encuesta

* RESULTADO:

Resultado obtenido (0 a 1, donde 0 es el menor resultado e indica el grado confiabilidad del entrevistado)

* TITULO\_CUESTIONARIO:

Nombre del cuestionario

Después de un estudio de la calidad de los datos se obtuvieron algunos hallazgos que se detallan en las siguientes líneas.

El análisis del atributo RESULTADO lo podemos ver a continuación, observando que tenemos valores 0, y el primer cuartil su valor es superior al 75% del rango del atributo, por lo que podríamos considerar que es un valor posiblemente erróneo y por lo tanto un registro inválido, o considerar el posible error en la migración y consolidación de información.

|  |  |
| --- | --- |
|  | **RESULTADO** |
| **count** | 32797.000000 |
| **mean** | 0.797175 |
| **std** | 0.275862 |
| **min** | 0.000000 |
| **25%** | 0.765300 |
| **50%** | 0.900000 |
| **75%** | 0.970000 |
| **max** | 1.000000 |

Se identifican 1743 registros con RESULTADO = 0, no es posible considerarla como un error de captura, pero podría corresponder a un error en la integración o exportación de los datos al generar el archivo.

El archivo consta de un volumen de 32767 registros y el rango de fechas en el campo de FEC\_EJECUCION es de 1995 a 2014, pero al revisar el detalle tenemos que de 1995 existe solo un registro, del año 2013 son 18 registros, existe un lapso de 17 años sin información (1996-2012), y de los años 1995 y 2013 realmente son muestras de registro muy pequeñas para tomar alguna decisión, y podrían contaminar el resto de la información, por lo que se deberá revisar con el cliente la exclusión de estos registros así como también la existencia de un registro sin fecha, en la siguiente tabla podrán identificarse en amarillo las observaciones señaladas. La misma información se encuentra agrupada por:

* Año
* Año y oficina
* Año y provincia



La siguiente tabla muestra la cantidad de registros por OFICINA y por PROVINCIA:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **OFICINA** | |  | **PROVINCIA** | | | | |
| **Oficina** | **Cantidad** |  | **Provincia** | **Cantidad** |  | **Provincia** | **Cantidad** |
| 911 | 9221 |  | MADRID | 6663 |  | VALLADOLID | 308 |
| 901 | 7915 |  | BARCELONA | 5438 |  | JAEN | 298 |
| 915 | 4469 |  | VALENCIA | 1879 |  | LEON | 280 |
| 917 | 2392 |  | ALICANTE | 1319 |  | CANTABRIA | 270 |
| 922 | 2171 |  | SEVILLA | 988 |  | CIUDAD REAL | 251 |
| 910 | 1970 |  | ASTURIAS | 941 |  | HUELVA | 238 |
| 905 | 1554 |  | A CORUÑA | 928 |  | ALBACETE | 233 |
| 908 | 1314 |  | TARRAGONA | 874 |  | TOLEDO | 230 |
| 919 | 710 |  | MALAGA | 859 |  | ORENSE | 221 |
| 921 | 640 |  | MURCIA | 793 |  | LUGO | 211 |
| SGS | 297 |  | PALMAS, LAS | 716 |  | SALAMANCA | 171 |
| 999 | 91 |  | SANTA CRUZ DE TENERIFE | 648 |  | ALAVA | 162 |
| 959 | 53 |  | CADIZ | 624 |  | CACERES | 140 |
|  |  |  | PONTEVEDRA | 607 |  | BURGOS | 128 |
|  |  |  | VIZCAYA | 603 |  | RIOJA, LA | 128 |
|  |  |  | GIRONA | 596 |  | HUESCA | 101 |
|  |  |  | BALEARES | 550 |  | SEGOVIA | 92 |
|  |  |  | LLEIDA | 487 |  | AVILA | 85 |
|  |  |  | ZARAGOZA | 467 |  | CUENCA | 80 |
|  |  |  | CORDOBA | 431 |  | PALENCIA | 77 |
|  |  |  | GRANADA | 397 |  | ZAMORA | 70 |
|  |  |  | CASTELLON | 362 |  | GUADALAJARA | 70 |
|  |  |  | EXTRANJERO | 358 |  | TERUEL | 46 |
|  |  |  | GUIPUZCOA | 353 |  | SORIA | 33 |
|  |  |  | ALMERIA | 317 |  | MELILLA | 31 |
|  |  |  | BADAJOZ | 311 |  | CEUTA | 25 |
|  |  |  | NAVARRA | 309 |  |  |  |

En la siguiente tabla hemos segmentado la información por OFICINA, se puede identificar la cardinalidad de cada atributo, así podemos observar que el atributo de COD\_LOC y NOMBRE\_LOC (dos atributos que aparentemente estarían relacionados uno a uno) para el caso de la oficina 911 tiene una relación prácticamente de 2 a 1, por lo que la inconsistencia es demasiado alta para considerar que existe relación entre ambos datos (se deberá aclarar con el cliente la interpretación correcta de estos datos, así como el valor que adiciona al análisis por realizar, ya que al ver del detalle de algunos registros parecieran ser datos de señalización o ubicación del cliente). Se observa también que las oficinas 901 y 911 concentran más del 50% de los registros.



Adicionalmente, la oficina 911 concentra 1671 registros de los 1743 con RESULTADO igual a CERO, esto equivale al 18.1% de sus registros, no tiene proporción si se considera que ninguna de las otras oficinas alcanza el 1%, por lo que se deberá aclarar con el cliente tal situación y quizá proceder a eliminar todos los registros con RESULTADO igual a CERO o repetir el proceso de exportación de datos de esa OFICINA. A su vez, la oficina 915 tiene menos de la mitad de los cuestionarios que las oficinas 911 y 901, pero tiene prácticamente la misma cantidad de cuestionarios con RESULTADO igual a UNO.

La oficina SGS es la única donde la cantidad de TITULO\_CUESTIONARIO es superior a la cantidad de COD\_PROY, también la codificación de éstos es diferente al resto de las oficinas, donde la codificación es ####\_###.

En la tabla siguiente, la información esta segmentada por PROVINCIA, se observa que BARCELONA y MADRID concentran un tercio de las evaluaciones, MADRID tiene el 25% de sus cuestionarios con RESULTADO igual a CERO y en contra partida, las provincias de BADAJOZ y HUELVA tienen más de 35% de sus cuestionarios con RESULTADO igual a UNO, que para el caso de MADRID y BARCELONA tales valores son el 11.6% y el 15.4% respectivamente.

Algunas provincias tienen Códigos Postales fuera de sus respectivos rangos, por lo que se sugiere revisar estos datos, así como también considerar que puede ser un dato irrelevante para el análisis por realizar.

La inconsistencia entre COD\_LOC y NOMBRE\_LOC se puede también observar.



Los datos de Población requieren aplicar un trabajo de limpieza, dado que existe duplicidad por falta de estandarización en el registro, en la primera tabla de este documento, se muestra que existen 1654 valores únicos para el atributo población, de acuerdo con la tabla siguiente, podemos observar algunos errores en la captura que provoca duplicidad de información:



Algo similar ocurre con el TITULO\_CUESTIONARIO, COD\_LOC y NOMBRE\_LOC, donde los registros no parecen tener un significado particular, sino textos libres de registro.

Podemos identificar tres valores importantes, OFICINA, PROVINCIA y COD\_PROY como posibles dimensiones para el análisis de la información dada la cardinalidad del atributo, la distribución de la compañía, la geografía del país y la interpretación que podemos hacer de los datos, y el atributo RESULTADO como el dato a medir.

1. **Análisis funcional y diagrama de arquitectura de flujo de datos.**

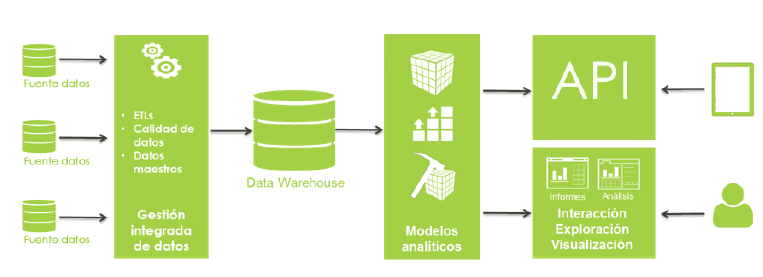
Realizaremos la carga del archivo a una base de datos relacional, realizaremos procesos de transformación y limpieza de datos (conversión a mayúsculas, abreviaturas, palabras con acento, diéresis, estandarizar el formato del código de proyecto, etcétera), posterior a ello se hará la extracción de los datos que corresponderán a las dimensiones que servirán para el análisis de la información.

Considerando el tipo de proceso será un DATA MART independiente, no volátil, la base de datos no realizará procesamiento de transacciones, recuperación, ni mecanismos de control de concurrencia más allá de las actualizaciones y consultas mismas que requerirán los procesos de inteligencia de negocio señalados en este documento.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| **ARCHIVO** |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Utilizaremos PENTAHO para los procesos de ETL, definición del cubo de datos y el análisis de información.

1. **¿Qué arquitectura de referencia usaría? Justifique la respuesta.**



Realizaremos un sistema con tres niveles (datos, aplicación y presentación) y crearemos un DATA MART considerando nuestra fuente datos como un repositorio ya existente y del cual realizaremos la extracción de información, lo cual nos dará la posibilidad de incrementar la funcionalidad con relativa facilidad en caso de ser esto solicitado.

En este primer proyecto tendremos un subconjunto de los datos departamentales, pero con posibilidades de incrementar los servicios integrando nuevos DATA MART o añadiendo otras estructuras al existente.

1. **¿Qué tecnología OLAP usaría? Justifique la respuesta.**

ROLAP, usaremos una base de datos relacional que nos permitirá un excelente tiempo de ejecución, se alinea perfectamente a la infraestructura y planes del cliente, requerirá un mínimo de capacitación y nos dará flexibilidad de crecer en más módulos BI de ser necesario en un futuro próximo.

1. **Si se utiliza ROLAP, ¿cuál de estos dos modelos se ajustaría mejor: el modelo en estrella o el de copo de nieve?**

ROLAP está basado en un sistema de base de datos relacional, ambos modelos (estrella o copo de nieve) funcionan adecuadamente en él, aunque hay señalar que un sistema estrella lleva comúnmente un proceso de DESNORMALIZACIÓN de las tablas que corresponderían a las DIMENSIONES, y uno de los fundamentos principales en una base de datos relacional es la NORMALIZACIÓN de sus tablas, aunque me gustaría señalar que la teoría de normalización surgió hace décadas (tales teorías son vigentes e incluso vitales al día de hoy para un sistema transaccional), donde el principal problema era la capacidad de los repositorios de datos, hoy en día la capacidad de un dispositivo de almacenamiento es miles de veces superior al existente en aquellos años, quizá hoy en día, el recurso más valioso que debemos considerar en todos nuestros diseños es el tiempo de procesamiento (esencial para un sistema de Business Intelligence) y que la información pueda ser entregada en la menor cantidad de tiempo posible.

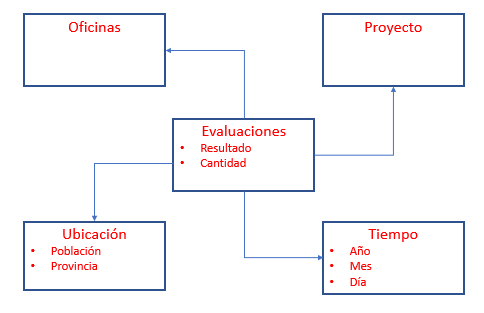
1. **Si se utiliza ROLAP, hay que identificar y justificar si existe algún proceso de desnormalización de información que se deba realizar.**

Utilizaremos ROLAP, de acuerdo con el análisis realizado y al diseño que se propone, tenemos una dependencia funcional con las que deberemos realizar un proceso de desnormalización, y esta es la que corresponde a la relación de POBLACIÓN – PROVINCIA. En un esquema normalizado tendríamos un catálogo con las provincias, y un catálogo de poblaciones, este último con una dependencia funcional al catálogo de provincias, en la desnormalización tendremos solamente un catálogo, que incluiría ambos datos (población y provincia).

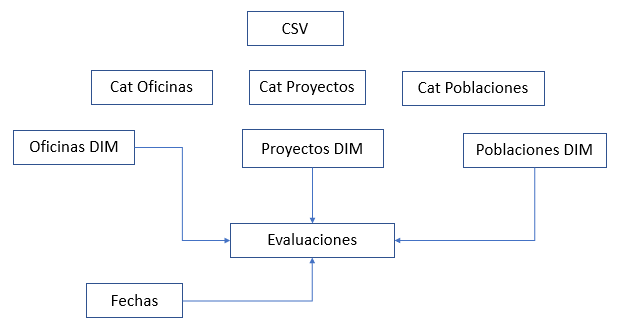
1. **Si se utiliza ROLAP, se debe incluir un diseño conceptual a modo explicativo junto con un diagrama.**

Nuestra propuesta es un esquema de estrella formado por 5 elementos:

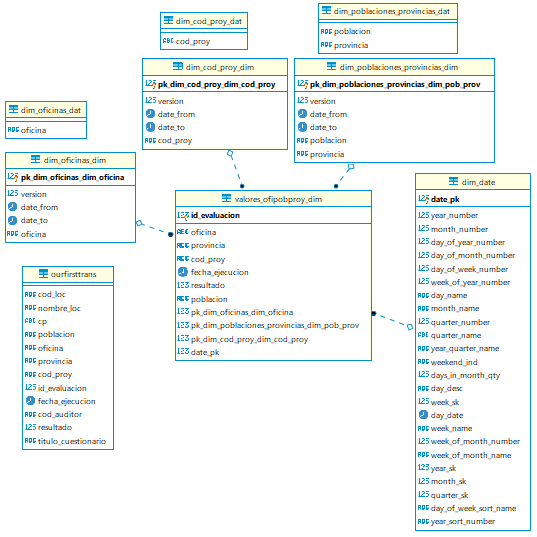
* Oficina (oficina o sucursal que realiza la evaluación)
* Proyecto (referencia de agrupación de los datos)
* Ubicación (ubicación del cliente evaluado)
  + Provincia
  + Población
* Tiempo (fecha de ejecución de la evaluación)
* Evaluaciones (cantidad y resultado)



1. **Si se utiliza ROLAP, se debe incluir un diseño modelo lógico.**



1. **Si se utiliza ROLAP, se debe incluir un diseño modelo físico**



1. **Realizar la implementación del proceso ETL para generar y poblar el modelo multidimensional diseñado en los apartados anteriores.**

|  |  |
| --- | --- |
|  |  |
| El proceso general de ETL para probar el modelo diseñado se muestra a continuación, el proceso tiene dos etapas, la primera de ella corresponde a la dimensión de TIEMPO, la cual es estática y solo es necesario generarla una vez, por motivos de eficiencia es separada del proceso general. Ambas etapas pueden observarse en los siguientes diagramas.  Proceso general de ETL *(\ourfirstjob \m05\_practfinal\_v2.kjb*): | |
|  | |
| Dimensión TIEMPO (*\ourfirstjob\dimensiones\DIM\_DATE.ktr*): |  |
|  | |

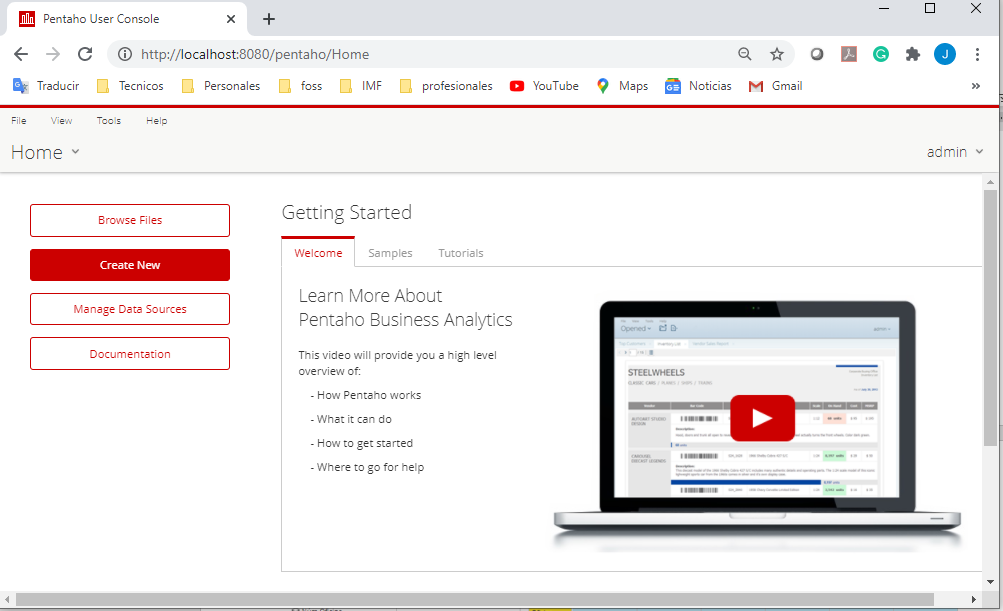
1. **Implementación de modelo multidimensional diseñado mediante los puntos anteriores. Se debe realizar con la herramienta Wizard facilitada y mostrada en vídeos anteriores.**

|  |
| --- |
|  |
| En las siguientes imágenes se muestra la configuración del cubo (de 4 dimensiones y sus jerarquías) mediante la aplicación Schema Workbench de PENTAHO y dos variables medibles (resultado y evaluaciones):   * + - * Oficina       * Provincia       * Proyectos       * Tiempo (año, mes día) |
|  |
|  |
|  |

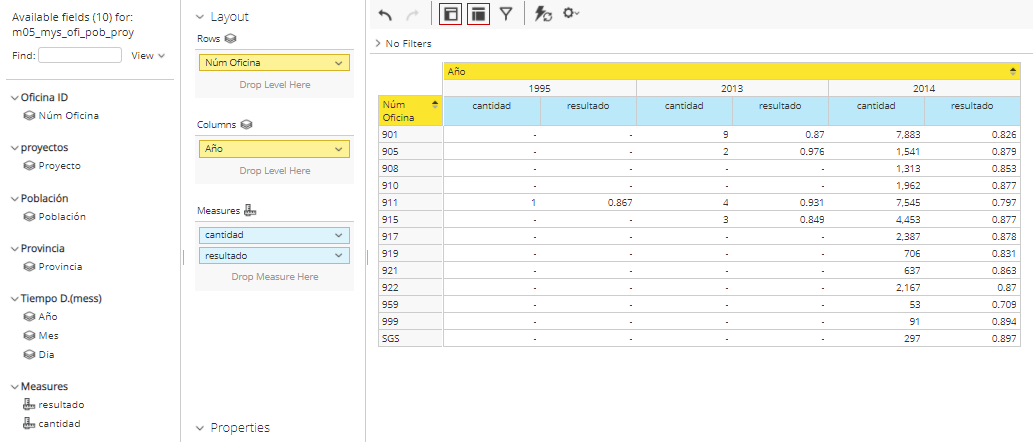
El detalle de la configuración del cubo se encuentra el anexo A de la documentación

1. **Análisis de modelo. Se solicita realizar, al menos, un análisis, haciendo uso de un modelo multidimensional que refleje alguna situación relevante de ser explicada y comentada. Para ello, se hará uso de los visores OLAP disponibles en la MV.**

A continuación, utilizando Analyzer de Pentaho (9.0) se muestran algunas consultas realizadas



En esta primera consulta tenemos la relación de oficinas y por cada año, la cantidad de evaluaciones registradas y el resultado obtenido



|  |  |
| --- | --- |
| Podemos definir una escala cromática sobre la variable a medir, para ellos seleccionamos el menú en la variable | Por año podemos observar cromáticamente en escala de verde-amarillo-rojo los valores para cada oficina, siendo verdes los mayores puntajes o cantidades a tonos rojos, para los menores valores. De esta forma rápidamente se puede apreciar que la mayor cantidad de encuestas fueron en el año 2014, para las oficinas 901 y 911, y que el peor resultado obtenido para el año 2014 fue 0.709 de la oficina 959 |
|  | La relación de oficinas, con la cantidad de encuestas realizadas, y el resultado (promedio) obtenido.  Es posible ordenar los resultados con clic en la columna.  Aquí se muestran ordenadas por el volumen de encuestas realizadas (izquierda) y por el promedio obtenido (derecha) |
|  | En esta imagen tenemos los datos obtenidos de las poblaciones ABADIÑO y ABRERA, segmentando la información por Oficina, Año y Mes, agrupada por Provincia y Población  Podemos ver que solo las oficinas 901 y 905 han participado en el año 2014 |
|  | Aquí se puede revisar el detalle de las oficinas 911 y 915 por año y mes, con los valores de encuestas realizadas y el resultado (promedio) obtenido incluyendo los totales de las dos oficinas filtradas |
|  | Configuración para obtener el resultado por PROYECTO y por OFICINA, aplicaremos también la escala cromática  Incluiremos también el promedio obtenido por proyecto |
|  | |
|  |  |
|  | Si definimos la consulta, y filtramos por el año 2014 |
| Cantidad de evaluaciones por oficina y provincia del año 2014 |  |
|  | Podemos cambiar al tipo de gráfica de Pie, oprimiendo el botón de la esquina superior |
|  | Aquí mostramos, por año, la cantidad de encuestas por cada oficina |
|  | Ahora aplicamos un filtro para mostrar solamente el año 2014 |
|  |  |
|  | Esta sería una gráfica de cada provincia, mostrando que oficinas y el porcentaje de encuestas de cada una. |
|  | Aplicando filtro a la provincia Navarra podemos observar lo que corresponde a cada oficina |
|  | Y cambiando el tipo de gráfica a Barras |
|  | Promedio obtenido por cada año, siendo el 2014 con el promedio mas bajo (0.842) |
|  | Y al dar 2-clic en cualquiera de las barras obtenemos el desglose por mes, en este caso el año 2104 |
|  | Las provincias con el mejor resultado obtenido durante el año 2014 |