**Series de tiempo, Agregación & Filtro**

## **Trabajando con Extractos de Datos en Tableau**

### **Vistas muy *pivotadas* vs vistas poco *pivotadas***

Compare las siguientes tablas, ambos contienen la misma información y reflejan series de tiempo.

Tabla

Descripción generada automáticamente  
vista *muy pivotada*

*Tabla

Descripción generada automáticamente*vista *poco pivotada.*

Es evidente que, para mayor facilidad de lectura, una tabla está mejor estructurada que otra. La tabla *muy pivotada* es poco eficiente, dificulta la rapidez con la que uno pueda conseguir e interpretar la información ahí plasmada; pues contiene, de forma excesiva, valores duplicados (lo cual molesta mucho a simple vista).

Curiosamente los **valores duplicados** son uno de los tantos tipos de **datos sucios (dirty data).**

Por otro lado, tenemos la tabla *poco pivotada* queevita en lo posible la cantidad excesiva de datos duplicados y, de hecho, lo logra; esto al permitir que los datos se clasifiquen y se enlisten en términos de su variable categórica *Age* y en términos de su variable **temporal** *Period*, organiza la información no sólo por columnas (campo ***Age***), sino también por filas (campo ***Period***): haciendo más eficiente la presentación de los datos de su tabla.

Adicionalmente, se filtra la información por género; es decir, por tabla sólo se ven los datos de un solo género, sea ***Men*** o ***Women;*** es quien tiene control sobre la hoja de cálculo el que decide qué datos ver, sea hombres o mujeres, al filtrarlos.

#### **¿Qué prefieren las maquinas?**

Aunque parezca sorpresivo, las maquinas u ordenadores prefieren lo contrario; es decir, trabajan mejor con las vistas *muy pivotadas* porque encapsulan, sin filtro alguno, toda la información de las tablas*.* En este tipo de casos puntuales es más fácil para Tableau, por ejemplo, importar este tipo de datos para su posterior análisis (le resulta más eficiente este tipo de formatos porque los estructuran y los preparan mejor). La estructuración y la preparación de datos son procesos muy importantes en el proceso completo de análisis de datos.

### **Cómo trabajar con extractos de datos en Tableau**

Los extractos son copias de datos que se hacen a partir de una fuente de datos importada en Tableau. Es la manera en cómo uno se conecta a una fuente de datos sin que la conexión sea en tiempo real; es decir, que la copia de datos generada en la importación, que resulta ser exacta hasta ese entonces, no pueda sufrir de alteraciones si la fuente de datos original se modifica inmediatamente después de la extracción hecha. En resumen, los datos van a ser extraídos en un fichero separado, siendo guardados en formato *Tableau* y dentro de Tableau propiamente.

Entonces, existen dos formas de conectarnos a los datos, por medio de una conexión en tiempo real o creando un *Extracto* de los datos.

#### **¿Cómo se crea un Extracto?**

Desde la hoja de trabajo que estás utilizando, para un determinado fichero pasado (en este caso por ejemplo hemos importado la fuente de datos ‘Long-Term-Unemployment-Statistics.xlsx’), hacemos clic derecho sobre lo siguiente:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente *Selecciona la opción* ***‘Extraer datos…’***

*Aviso de interés: Es preciso mencionar que esta opción sólo está habilitada para Tableau Desktop.*

Al hacer clic sobre ***‘Extraer datos…’,*** aparecerá la siguiente ventana emergente:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Aquí podemos determinar si hacer una copia completa de los datos extraídos, es decir, hacer una copia de todos los datos del fichero o, por el contrario, hacer una copia solamente un subconjunto de esos datos.

Por defecto, al estar seleccionada la opción **‘*Todas las filas*’**, se hará una copia total de los datos al hacer clic en **Extracción**. Posteriormente, guarda su extracción en el directorio de su selección.

Ahora, usted puede decidir si desea mantenerse trabajando con su extracto o, por el contrario, continuar trabajando con la fuente de datos original que se actualiza en tiempo real; todo a un solo clic. Veamos.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Si quita el *checkmark* (marca de verificación) de **‘*Usar extracción*’**, Tableau interpretará inmediatamente que desea volver a trabajar con la fuente de datos original; la cual puede sufrir modificaciones y alterar, de paso, la información que tiene importada en Tableau, pues, ya no estará trabajando con el fichero separado que se creó al generar un Extracto con formato *Tableau,* sino, con el fichero original de donde extrajo toda su información.

Conclusión: Según su necesidad, usted decidirá si es más conveniente trabajar con una conexión en tiempo real o con un Extracto.

## **Trabajando con Series de Tiempo**

En esta sección estaremos trabajando con el dataset 'Long-Term-Unemployment-Statistics' que trata acerca de las estadísticas del desempleo en EE. UU. a largo plazo. Nos interesa ver cómo evoluciona la variable o campo del desempleo, *Unemployed*,a través del tiempo. Es preciso recordar que, como se trata del análisis de una sola variable a través del tiempo, lo ideal sería hacer uso de visualizaciones lineales (gráficos lineales).

Dicho lo anterior, entonces, veremos cómo visualizar y cómo varían las visualizaciones del campo *Unemployed* a lo largo del tiempo: se buscarán tendencias, se agruparán datos por categorías y mucho más.

En esencia, trataremos de poner en práctica lo que significa trabajar con datos en series temporales desde Tableau.

Empecemos. Partamos de que ya hemos importado en Tableau la fuente de datos 'Long-Term-Unemployment-Statistics', que resulta ser un archivo de tipo *Excel.* Nosotros podríamos, desde nuestra Worksheet, ver los datos crudos de dicha fuente de datos antes de visualizarlos propiamente; digamos, para tener una idea previa de lo que se verá reflejado en la posterior visualización ofrecida por Tableau. Tal que así:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Posteriormente, verá la siguiente ventana emergente con todos sus datos de forma estructurada, casi como si los viera directamente desde su hoja de cálculo. Observe:

Tabla

Descripción generada automáticamente

### **Al grano: Gráfico de Series de Tiempo**

Como se comentó anteriormente, nos interesa ver cómo evoluciona la variable o el campo del desempleo, *Unemployed*,a través del tiempo.

Es preciso recordar que los campos de tipo *temporal*,los que manejan fechas, que son un tipo de dato cuantitativo, no cualitativo (en nuestro caso puntual *Period*),al arrastrarlos al *dashboard* deben ordenarse por Columnas. Por el contrario, como usted ya conoce, los campos cuantitativos como *Unemployed* se ordenan por Filas; es decir, de estos campos es que se definen los ejes, ***x*** e ***y***, las ***columnas*** y ***filas***, para la visualización de nuestro gráfico y el estudio de nuestra variable de interés en el tiempo: ***Unemployed***. Veamos esto inmediatamente en un gráfico.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

#### **Agregación vs Categorización (Granularidad)**

Si se da cuenta, el campo *Period* se está presentando en términos anuales (fechas presentadas en términos de *años,* y no *meses*); es decir, se ha acumulado (y reducido) toda la información de cada 12 meses, en términos de años, para tener una visualización más *digerible* y resumida. A este ejercicio de agrupar subconjuntos relacionados (12 meses) en su conjunto principal (año) se le conoce técnicamente como: **Agregación.** La **Agregación** es justamente lo contrario a una **Categorización o Granularidad.** Una **Categorización** haría divisiones en la variable de interés.

Ahora, si se percata, esta agregación se ha hecho independientemente de la ***edad*** y ***genero (Age & Gender, respectivamente)***; pues, nuestro gráfico inicial no está tomando en cuenta las distinciones de los valores de dichos campos, simplemente presenta la información de forma generalizada **(suma la cantidad total de todos los desempleados, año a año, sin distinguir género o edad de la persona desempleada en cuestión)**.

Lo anterior está interesante porque podemos ver, con más facilidad, las tendencias que se crean a partir de la visualización reflejada. Puede notar como, desde el año 2008, el desempleo en los Estados Unidos se dispara; a partir de ahí usted puede reflexionar, pensar y cuestionarse porqué o cuáles fueron las razones macroeconómicas que condujeron a eso, ¿Esto acaso no se debió a la crisis inmobiliaria del año 2008? Ya puede ir extrayendo un par de *Insights* con ayuda de su visualización.

#### **Granularidad**

Ahora vamos a *Granular,* a *categorizar*,los datos ofrecidos por esta visualización inicial; es decir, ahora sí vamos a llevar a la visualización a un nivel de detalles superior. Veamos si podemos categorizar en meses la evaluación del desempleo a lo largo de toda la línea de tiempo del dataset (Granular el tiempo). Debe hacer clic sobre el campo temporal ya arrastrado, *Period,* y ver el siguiente menú despegable.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Dentro del menú despegable hay dos secciones responsables propiamente del tratamiento del tiempo, puede intuir cuáles son.

En la primera sección, podrá determinar con qué marco temporal representar su serie de tiempo. Sin embargo, hagamos una pausa… Realmente si bajamos a meses, por ejemplo, notará algo particular: dentro de este proceso de *Granularidad* se dará un proceso de *Agregación* indeseado. Explico. Si bien los datos serán presentados en términos de meses, éstos se *agregarán,* se sumarán de tal manera que se mostrará, para cada uno de los meses, la totalidad de todos los desempleados que hay dentro de toda la línea de tiempo evaluada en el *dataset*; es decir, se cuentan todos los desempleados que hay para todo el *dataset*, sin discriminar edad ni género, peroclasificándolos por mes… Veamos cómo se vería este gráfico.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Esto es muy poco práctico visualmente. De hecho, más allá de lo poco práctico que se ve visualmente, no tiene ningún sentido sumar la cantidad de desempleados que hay, por ejemplo, en el año 2013 con los desempleados del año 2014 y 2015; peor, no tiene ningún sentido acumular o sumar todos los desempleados que hay en toda la cronología del *dataset* y solo distinguirlos por un determinado marco temporal y el marco temporal *mensual* no es la excepción, **¿Qué clase de análisis vamos a realizar si no podemos evaluar la variación del desempleo año a año, es decir, si no podemos evaluar la variación en las tendencias a lo largo del tiempo?**

La anterior ‘falla’ acontece debido a que Tableau no está interpretando al campo, en este caso *Period*, como campo de tipo *temporal*, sino como un campo de tipo cualitativo o una *dimensión*; por lo que a Tableau, en un principio, no le interesa evaluar tendencias, variaciones, ni líneas de tiempo con un determinado campo que supone ser *temporal*, pues, no está siendo tratado o interpretado como tal.

#### **Granularidad Real**

Para lograr que un campo de tipo *Temporal* sea tratado como tal, en este caso *Periodo*, a la hora de hacer clic sobre el campo temporal ya arrastrado, y ver su correspondiente menú despegable; debe seleccionar el marco temporal, con el que pretende visualizar sus datos, **desde la 2da sección responsable del tratamiento del tiempo.** Acá:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente, *observe ahora sí un gráfico correcto con temporalidad* ***Mes****:*

Gráfico, Gráfico de líneas

Descripción generada automáticamente

*Más de cerca…*

*Gráfico, Gráfico de líneas

Descripción generada automáticamente*

De hecho, como se explicó antes, los campos temporales son un tipo de campo cuantitativo (medida), ¿cómo se da cuenta de que en efecto está delante de una medida? Observe ahora el color del campo temporal *Period* ya arrastrado, es **de color verde, indicativo de ser una medida**; contrario a **los campos cualitativos o *dimensiones*, son de color** azul (de hecho, si se percató bien, antes el campo *Period* era de color Azul por no ser una *medida* realmente). ***Esta última visualización, en definitiva, ¡sí que es una serie temporal!***

#### **Marcadores**

Ahora, si queremos hacer uso de algún recurso que evidencie, de forma física, la granularidad a nivel de cualquier marco temporal seleccionado (***mes*** para nuestro caso); es decir, que Tableau tome nuestra última visualización y le pase un *marcador*, visible a los ojos, que destaque los meses de nuestra serie de tiempo… podemos hacer uso de **marcadores.**

Luego de tener diseñado su serie de tiempo, váyase a la marca ‘***color’*** y haga clic sobre ella; posteriormente, diríjase a la sección de ‘***marcadores’*** (al final) y determine que diseño usar. Tal que así, por ejemplo:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Los marcadores, como puede ver, facilitan la lectura de nuestra granularidad; es decir, para este caso puntual, nos facilita la lectura del nivel de desempleo en los Estados Unidos, mes a mes, en nuestra serie de tiempo (originalmente la serie de tiempo va desde el año 2005 hasta el año 2015).

**Ahora, un detalle importante que debe recordar.** De momento no estamos haciendo distinciones de la cantidad de desempleados, mes a mes, según género o edad, no. Estamos *agregando* hasta ahora, sumando, todos los desempleados sin distinguir género o edad, mes a mes, desde el año 2005 hasta el año 2015. Posteriormente, en efecto, sí vamos a *categorizar* o *granular* la cantidad de desempleados, ya no solo por mes; sino, también, por género y edad. En el siguiente apartado profundizaremos esto. De momento hay una granularidad por *mes,* ahora querríamos llevar las cosas a un *nivel de detalle* más elevado; ahora vamos a granular no sólo por mes, sino también por género y edad.

**Dato curioso:** En el caso extremo, si quisiéramos saber el *nivel de detalle* más simple posible (nivel global de los datos)*,* lo que personalmente llamaría *la gran antítesis de la categorización/granularidad o la gran tesis de la agregación*, podríamos simplemente representar la suma total de todos los desempleados que ha habido en EE. UU. desde el año 2005 hasta el año 2015, sin categorizar absolutamente nada. Esto sería la representación de un solo punto, un solo valor, en nuestra visualización:

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

#### **Conclusiones de interés**

Cuando hablamos de **granularidad del dato** nos referimos al nivel de detalle con el que se almacena y se analiza una información. Por ejemplo, una empresa que se dedique a la venta de víveres puede registrar de forma diaria lo que vende y, por otro lado, registrar mensualmente información relativa a pagos de sueldos.

La posibilidad de analizar la información con mayor profundidad va a depender del nivel de detalle con el que se almacenen los datos.

Por otro lado, también es importante mencionar que la **granularidad del dato** trabaja en conjunto con la agregación. Si lo extrapolamos a las bases de datos, se trata de un cálculo que se hace a varias filas y no produce un único resultado.

Lo relevante para las empresas es el **nivel de granularidad de los datos** de sus negocios para poder tomar decisiones adecuadas. Si se agregan los datos de todas las ventas, se podrán analizar las ventas en el último año de un producto en concreto. Por tanto, se pueden hacer consultas y tener una repuesta de los análisis de datos de forma ágil. **Tomado de**: [*https://www.upicus.com/es/granularidad-del-dato/#:~:text=Cuando%20hablamos%20de%20granularidad%20del,relativa%20a%20pagos%20de%20sueldos*](https://www.upicus.com/es/granularidad-del-dato/#:~:text=Cuando%20hablamos%20de%20granularidad%20del,relativa%20a%20pagos%20de%20sueldos)*.*

## **Agregación, Granularidad & Nivel de detalle: A fondo.**

### **Medidas & Agregación vs Dimensiones & Granularidad**

En teoría, o al menos en lo fundamental, usted ya tiene una noción clara del significado de estos conceptos. A partir de lo que ya hemos explicado relacionado a este tema, por el anterior apartado, sepa de entrada lo siguiente: **Los campos tipo *medidas*, en general, van a ser *Agregados* o son los capaces de generar *Agregación*; mientras, los campos tipo *dimensiones* o *temporales* son las capaces de especificar el nivel de *Granularidad.***

De hecho, nuestro caso de estudio actual evidencia la anterior relación. Note que su campo *Period*, que es de tipo *temporal*, es quién realmente puede (entre otros) *granular* o *categorizar* la serie de tiempo (también lo pueden hacer los campos *Gender* & *Age,* que son *dimensiones*); en cambio, nuestro campo *Unemployed,* que es una *medida,* es quien va a generar la *Agregación* en nuestra serie de tiempo: Si usted solo arrastra el campo *Unemployed* al *dashboard* se dará cuenta que, por defecto, la información de su visualización será presentada de la manera más *global* posible, con el nivel de detalle más simple; es decir, verá el valor de la suma de todos los desempleados que ha habido entre el 2005 al 2015, sin distinguir o categorizar absolutamente nada.

Resulta que, por defecto, las *medidas* al ser arrastradas al *Dashboard* **agregan.** Por ejemplo, el campo *Unemployed* al **sumar** a todos los desempleados, y no presentar cada registro de desempleo por separado (tal y como viene desde la fuente de datos), está ***agregando…*** Sin embargo, sepa que, si no desea agregar por defecto sus valores, por medio de sus *medidas*, también puede hacerlo. Usted puede suprimir su Agregación.

#### **Excluyendo la Agregación**

Lo que indica que, en efecto, las *medidas* por defecto generan *Agregación* es lo siguiente: Si usted se va a la pestaña de **Análisis,** de la barra de navegación principal de Tableau, verá que está preseleccionada la opción que dice: ‘***Agregar medidas’***; pues bien, es esta opción que, estando validada por defecto, hace validar también el hecho de que las *medidas* **agreguen** o generen **Agregación.**

**Si usted deselecciona la opción ‘Agregar medidas’**, notará que la agregación de su campo de tipo *medida* se perdería; presentando así la información de los valores o datos, de la *medida* en cuestión, tal cual están registrados en la fuente de datos importada. Probemos esto:   
Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

*Veamos el gráfico resultante…*

Gráfico, Histograma

Descripción generada automáticamente  
  
En principio, por medio de este tipo de gráfico automático, no se percata de que en efecto se ha perdido la *Agregación*; es decir, de momento no se percata, de la mejor manera visual, de que ahora tiene evidenciado cada uno de los registros de desempleo que guarda año a año, mes a mes, la fuente de datos original (nuestro archivo de Excel importado). En resumen, no se percata de la mejor manera la visualización explicita de cada registro de desempleo que hay en la fuente de datos original; esto si se toma en cuenta que cada uno de los registros de desempleo están organizados por rangos de edades y géneros sexuales.

Sin embargo, podemos solucionar este detalle. Cambiemos el gráfico lineal a uno de círculos para que perciba mejor, por medio de puntos distinguibles, que ahora NO hay agregación; es decir, ahora podrá visualizar la totalidad de cada uno de los registros de desempleo… tal cual los registró su fuente de datos original (que toma en cuenta los registros de desempleo según rango de edad y género; aunque por ahora no los distinguiremos visualmente, simplemente se mostrará la totalidad de los registros sin agregación pero tampoco sin categorizarlos -los registros de los desempleados- por género ni por rango de edades). Observe.

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

*Cada punto representa un registro de desempleados para la fuente de datos importada, ¡ahora sí se percata visualmente la Anulación de la Agregación!* **Anular la Agregación** es conveniente en el caso de que tenga un gran interés de seguir granulando sus datos; es decir, categorizándolos… justamente como lo haremos, ahora sí, con los *rangos de edad* y *genero* (*dimensiones*) de la totalidad de los desempleados registrados desde el año 2005 al año 2015 en EE. UU.

#### **Granulando nuestras dimensiones**

En efecto, si quisiéramos distinguir visualmente cada uno de los registros del desempleo por *rango de edades* y *género*, como ya se comentó, también se puede hacer.

¿Qué mejor recurso inicial que la marca ***Color***para hacer visuales las distinciones que hace una dimensión? Empecemos por distinguir con la marca *Color* la dimensión *Gender* en los registros que hay del desempleo*.*Es decir, ahora haremos visual la categorización del *género* en la totalidad de los desempleados que hay en la serie de tiempo.

Me explico, ahora verá (visualmente) cuántas mujeres desempleadas hubo desde el año 2005 hasta el año 2015, mes a mes, y lo podemos comparar con la cantidad de desempleados *hombres* para ese mismo intervalo de tiempo. Esto sin distinguir aún, visualmente, los rangos de edad. **Arrastremos** ***Gender* a la marca *Color*.** Veamos.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Ahora nuestro nivel de granularidad comprende el periodo (Period) mensual y el género (Gender).

Sin embargo, si lo que se desea es llegar a un nivel de granularidad completa; es decir, distinguir o categorizar los niveles de desempleo también por rango de edades, que sería nuestra última *dimensión* por evaluar según nuestra fuente de datos original, podemos hacer uso de otra *marca* (No puede ser la marca *color* porque con los colores ya se distingue la *dimensión* del género sexual, es decir, los colores ya distinguen la sexualidad).

Una marca interesante para distinguir el rango de edades podría ser ***Tamaño.*** Es decir,dado un tamaño del circulo podríamos saber en qué rango de edades se ubican los desempleados.

**Arrastremos, entonces, la dimensión *Age* a la marca *Tamaños.*** El gráfico quedaría tal que así:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Ahora nuestro nivel de granularidad comprende el periodo (Period) mensual, el género (Gender) & el rango de edades (Age).

**¿Qué *Insights* podríamos extraer de nuestra última visualización? Reflexiones sobre esto.**

#### **Alterando la Agregación**

Si quisiéramos considerar de nuevo el efecto de la Agregación en nuestro gráfico, debe saber también que hay diferentes tipos de agregación que podemos fijar con nuestros campos de tipo *medida*; es decir, aquellos campos que realmente pueden *agregar*. En nuestro caso, como ya lo sabe, la *medida* con la cual se está *agregando* la visualización es: ***Unemployed.***

Para nuestra mayor comodidad, supongamos que retomamos nuestro punto de partida inicial:

“… Resulta que, por defecto, las *medidas* al ser arrastradas al *Dashboard* **agregan.** Por ejemplo, el campo *Unemployed* al **sumar** a todos los desempleados, y no presentar cada registro de desempleo por separado (tal y como viene desde la fuente de datos), está ***agregando…****”.* Es decir, supongamos que estamos de nuevo en el punto donde abordamos nuestra agregación inicial, la que se hace por defecto: la **suma** de los valores de la *medida* Unemployed. Omitimos cualquier tipo de granularidad, a excepción de la granularidad que comprende el periodo mensual.

Veamos el gráfico de nuevo.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Si quisiéramos cambiar el tipo de agregación; es decir, el cálculo que opera el campo de tipo *medida* en juego, lo que se conoce también como ***Operación de Agregado…*** Podemos hacerlo de la siguiente forma:

***Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente***

Observe todas las Medidas que hay, son el equivalente a los tipos de agregación existentes o las operaciones de Agregado que podría fijar: **Suma, Promedio, Medina, Recuento, etc.** Probemos **Promedio**, por ejemplo.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

En cuanto a tendencia el gráfico no cambia demasiado, pero sí que cambia en el Eje. Note que los valores más altos están en torno a los 500mil; mientras que, cuando en lugar de hacer el *promedio* hacemos la *suma,* los valores llegan hasta los 7millones. **Entonces, esta última visualización nos permite saber el promedio de los desempleados que hubo, mes a mes, y sin distinguir por género o rango de edad, para la totalidad de los registros encontrados en la fuente de datos original: “Long-Term-Unemployment-Statistics”.**