

1. PV.10 - CONEXIÓN DE CONJUNTOS DE DATOS

1.1. Objetivo:

Analizar dos conjuntos de datos al mismo tiempo combinándolos en Tableau.

1.2. Resultados del aprendizaje:

- Identificar datos comunes entre conjuntos de datos que permitan conectarlos.
- Generar un campo común que facilite la conexión por parte de software como Tableau.
- Analizar los datos de dos conjuntos de datos diferentes una vez combinados.

En este ejercicio, trabajarás con dos conjuntos de datos:

- Resultados de las elecciones presidenciales de 2012 por distrito del Congreso (435 filas, sólo la Cámara de Representantes) adaptados del sitio web Daily Kos. Proporciona el porcentaje de votos otorgados a Romney y a Obama en cada distrito del Congreso. También tiene un campo que enumera quién ganó ese distrito.
- Los perfiles demográficos de cada uno de los congresistas actuales (535 filas, Cámara de Representantes y Senado) del proyecto Measure of America, que forma parte del Social Science Research Council. El conjunto de datos incluye el partido político, el sexo, la raza y el nivel educativo del funcionario elegido (también hay otros datos).

Combinando estos conjuntos de datos, podemos averiguar si parece haber relaciones entre la demografía del representante elegido por el distrito y el voto de ese distrito en las elecciones presidenciales de 2012.

Hay que tener en cuenta que la correlación no siempre implica causalidad. Cuando vemos algo que parece una relación, no significa necesariamente que entendamos la causa, puede tratarse de una simple coincidencia. Pero sigue siendo interesante observar ...

1.3. Parte 0: Echar un vistazo a los conjuntos de datos

1. Abrir el archivo "2012 Presidential Election Results by District.xlsx" y mirar los datos. Veréis una entrada para cada distrito del Congreso (es decir, AZ-1, AZ-2, AZ-3, ...). Cada estado tiene al menos un distrito, dependiendo del tamaño de la población. Contiene el porcentaje de votos para Obama y Romney (no sumará el 100% porque siempre hay candidatos de terceros partidos y candidatos por distrito). También verá que el Estado y el Numero de Distrito están divididos en columnas separadas. Tenemos que hacer esto para poder hacer cosas interesantes de mapeo con Tableau.
2. Ahora, abrir el archivo "Portrait 113th Congress.xlsx" en Excel y mirar los datos. Aquí verás una lista de cada representante elegido y su información demográfica. Mirar la fila 10 (la primera fila de datos). Observar que District (IF HOUSE) es solo un número, en lugar de AL-1, como se representó en el archivo de resultados electorales. Estos formatos diferentes para el distrito harán imposible que Tableau conecte los datos más tarde: no podrá averiguar que "Alabama 1" es lo mismo que "AL-1". Así que tendremos que arreglar esto antes de hacer nuestro análisis.
3. Cerrar ambos archivos de Excel.

1.4. Parte 1: Crear un campo común para combinar datos

Como hemos dicho anteriormente, el conjunto de datos "2012 Presidential Election Results by District.xlsx" representa los distritos de una manera y "Portrait 113th Congress.xlsx" de otra como vemos aquí:

	A
1	District
2	AK-1
3	AL-1
4	AL-2
5	AL-3
6	AL-4
7	AL-5
8	AL-6

STATE	DISTRICT (IF HOUSE)
Alabama	1
Alabama	2
Alabama	3
Alabama	4
Alabama	5
Alabama	6

Necesitáis crear una columna de datos adicional en uno de los archivos que represente los distritos de la misma manera. Necesitamos una sola columna para hacer el cotejo, así que vamos a modificar el archivo "Portrait" añadiéndole una columna adicional con una sola etiqueta de distrito.

1. Abrir el archivo "Portrait 113th Congress.xlsx" en Excel.
2. Observar que hay una pestaña "StateLookup". En ella nos encontraremos las abreviaturas de todos los estados ordenadas por orden alfabético.
3. En primer lugar, en la hoja "Data", iremos a la Columna M y en la celda M9 pondremos "STATEABBR". Crear una fórmula con BUSCARV para obtener la abreviatura del estado y la colocaremos en M10.
4. En la columna N en la celda N9 pondremos "DISTRICTCODE" y crearemos una fórmula que nos dé el formato Abreviatura-número de distrito.
5. Una vez comprobado que funciona extrapolamos estas fórmulas al resto de celdas de las columnas M y N con datos de estado. Deberíamos tener algo como esto:

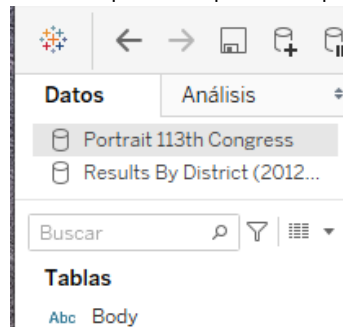
STATEABBR	DISTRICTCODE
AL	AL-1
AL	AL-2
AL	AL-3
AL	AL-4
AL	AL-5
AL	AL-6

6. Utilizaréis los datos de la columna DISTRICTCODE más adelante para conectar los dos libros de Excel, ya que ahora es exactamente igual que "District" en el otro archivo.
7. Guardar el archivo Excel con el nombre "Portrait 113th Congress-adaptado.xlsx".

1.5. Parte 2: Abrir los archivos de datos y conectarlos las fuentes de datos

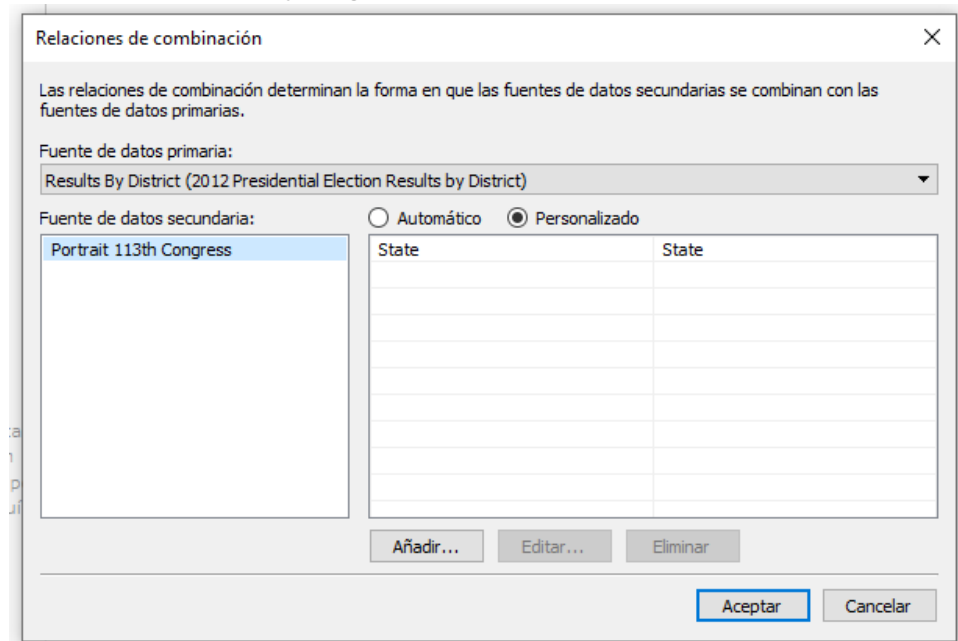
1. Iniciar Tableau y abrir el archivo "2012 Presidential Election Results by District.xlsx".
2. Arrastrar la hoja "Results by District" al espacio de trabajo.
3. Hacer clic en "Ir a la hoja de trabajo".
4. Ahora os conectareis al segundo archivo de datos. Ir en el menú a Datos y seleccione "Nueva fuente de datos".
5. Esta vez, abrir el archivo "Portrait 113th Congress-adaptado.xlsx".

6. Arrastrar la hoja de trabajo "Data" al espacio de trabajo y haga clic en "Ir a la hoja de trabajo".
7. Ahora veréis dos fuentes de datos en la parte superior izquierda de la ventana de Tableau:

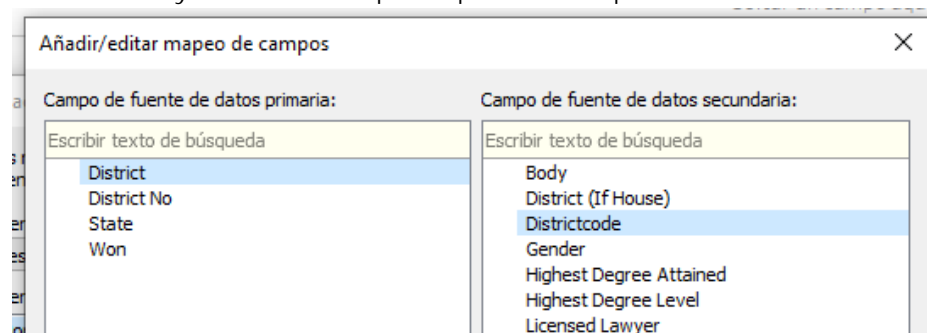


Habéis abierto ambos archivos, pero siguen sin estar conectados. Sabéis, sin embargo, que "DistrictCode" y "District" tienen el mismo formato. Podréis usar estos campos para conectar los datos de ambas fuentes en vuestro análisis.

8. Vaya al menú Datos y seleccione "Editar relaciones de combinación...".
9. Hay que asegurarse de que "Results by District" está como Fuente de datos primaria.
10. Seleccionar "Personalizado" y luego hacer clic en Añadir...

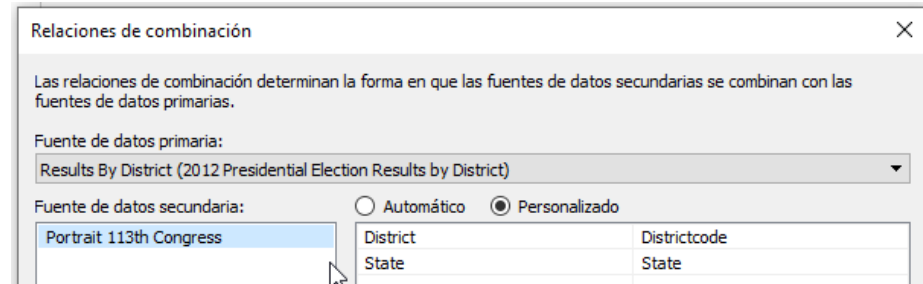


11. Seleccionar "District" y "DistrictCode" para que ambos queden resaltados.



A continuación, haga clic en Aceptar.

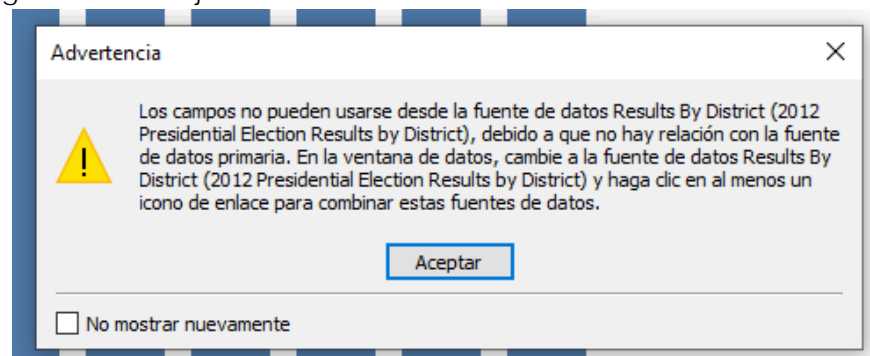
12. Volverá al diálogo anterior:



13. Eliminar la relación State-State y haga clic en Aceptar.

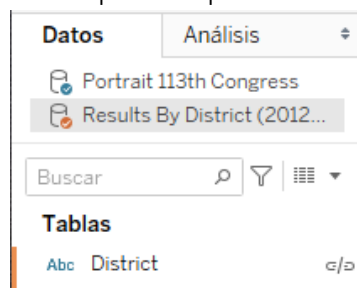
1.6. Parte 3: Crear un gráfico con los datos de ambas fuentes

1. Hacer clic en la fuente de datos "Data(Portrait 113th Congress)" en la parte superior izquierda.
2. Arrastrar la dimensión "Party" a Columnas.
3. Hacer clic en la fuente de datos "Results by District(2012 Presidential Election Results by District)".
4. Arrastrar la medida "Obama 2012" a Filas.
5. Veréis el siguiente mensaje:

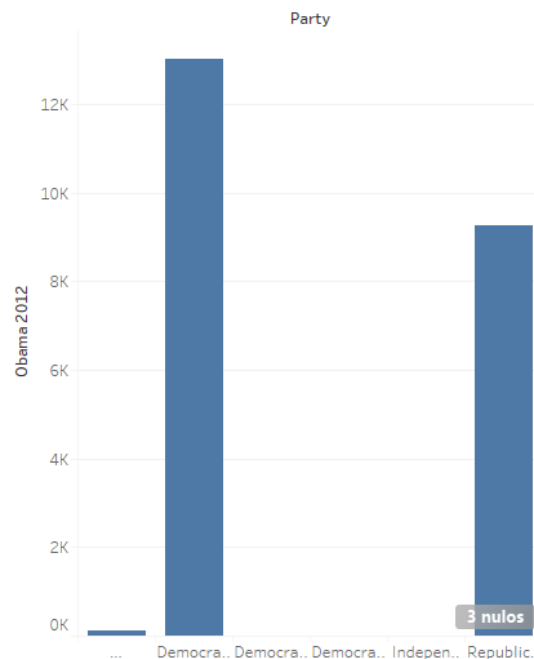


Hacer clic en Aceptar.

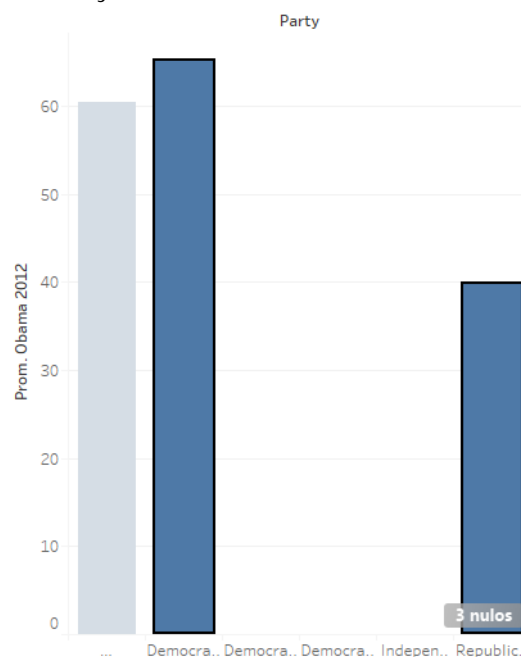
6. Ahora verá lo siguiente en la parte superior izquierda de su ventana de Tableau:



- Haga clic en el enlace roto en District. El enlace cambiará a un enlace naranja conectado y el gráfico tendrá este aspecto.

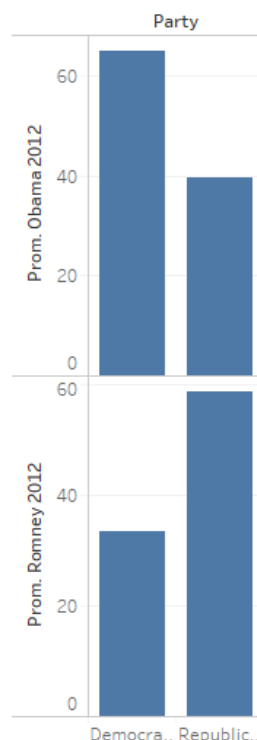


- Hacer clic con el botón derecho en SUMA(Obama 2012) en Filas y seleccionar Promedio.
- Mantener pulsada la tecla Ctrl y hacer clic en las barras demócrata y republicana.



- Pasar el ratón por encima de cualquiera de las barras resaltadas y seleccionar "Mantener solo".
- Arrastrar "Romney 2012" desde medidas y colocarlo junto a PROM(Obama 2012) en Filas.
- Hacer clic con el botón derecho en SUMA(Romney 2012) y cambiarlo a Promedio.

13. El resultado deberá ser como esto:



Con este gráfico sabemos que en los distritos del Congreso en los que el representante elegido es demócrata, Obama obtuvo una media del 65% de los votos frente al 33% de Romney. En los distritos en los que el representante elegido es republicano, Romney obtuvo una media del 59% de los votos frente al 40% de Obama.

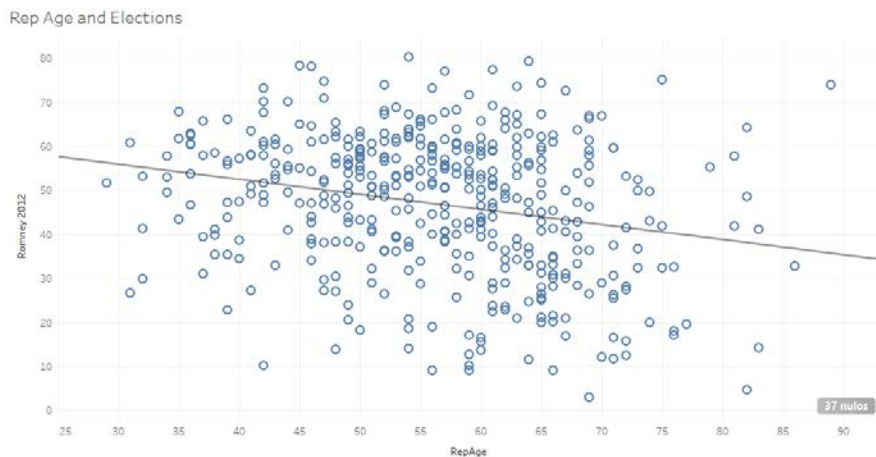
Esto fue posible al haber combinado los datos de los resultados electorales de la hoja de trabajo "2012 Presidential Election Results by District" con los datos de los partidos políticos de la hoja de trabajo "Portrait 113th Congress".

14. Nombrar la hoja como "Rep Party and Election Results" y guardar el trabajo.
15. Duplicar la hoja de trabajo y cambiar el nombre a "Rep Gender and Election Results".
16. Determinar si los distritos que eligen mujeres como representantes tienen más probabilidades de votar a Obama o a Romney.

1.7. Parte 4: Combinar un campo calculado en una fuente de datos con los datos originales de la otra

1. Crear una nueva hoja de trabajo. Nombrar la hoja "Rep Age and Election Results".
2. Hacer clic en la fuente de datos "Data(Portrait 113th Congress)".
Crear un campo calculado haciendo clic en Análisis>Crear Campo calculado. Llama al campo "RepAge" y utilice la fórmula $2012 - [\text{Year Of Birth}]$. Calcula la edad del representante restando el año de nacimiento al año actual (2012 porque es cuando se celebraron las elecciones).
3. Arrastrar RepAge de medidas a Columnas.
4. Hacer clic en la fuente de datos "Results By District(2012 Presidential Election Results by District)".
5. Arrastrar Romney 2012 de medidas a Filas. (Cuando veáis de nuevo el mensaje de advertencia hacer clic en Aceptar. A continuación, hacer clic en el enlace roto).
6. Hacer clic derecho en SUMA(RepAge) y seleccionar Dimensión. Hacer lo mismo para SUMA(Romney 2012).

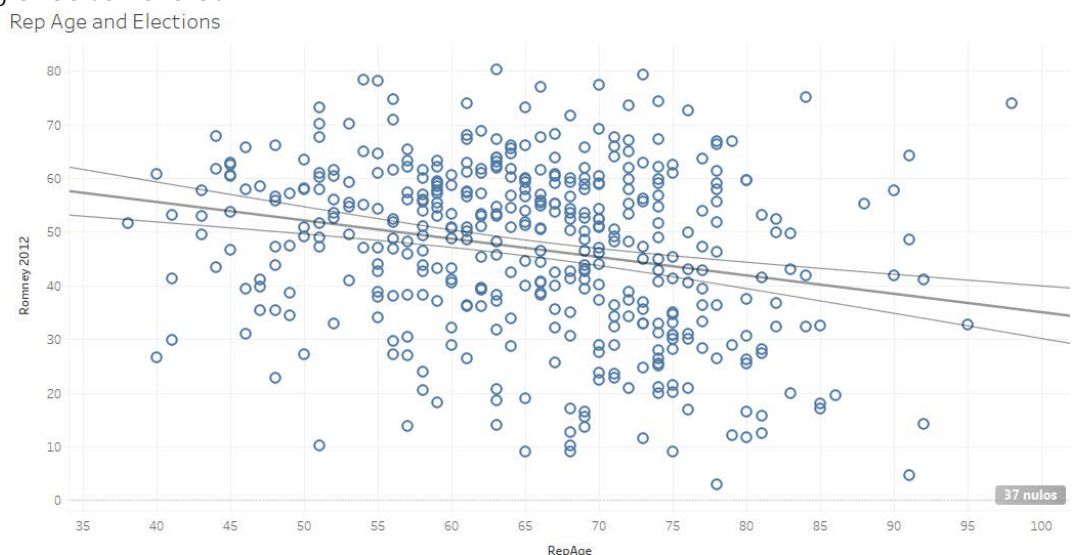
7. Hacer clic con el botón derecho dentro del gráfico de dispersión y hacer clic en “Líneas de tendencia>Mostrar líneas de tendencia”. Verás esto:



Esto implica una relación negativa entre la edad del diputado elegido y el hecho de que ese distrito haya votado a Romney.

Si volvemos a hacer clic derecho dentro del gráfico y seleccionamos “Líneas de tendencia>Editar todas las líneas de tendencia...” se abrirá una ventana con diversas opciones. Vamos a seleccionar la opción de Mostrar bandas de confianza.

El gráfico se vería así:



1.8. Entregable

Esta tarea debe realizarse de forma individual. Complete las instrucciones de este documento y envíe a Moodle:

- El **fichero Excel** "Portrait 113th Congress-adaptado.xlsx"
- El **fichero de Tableau** (extensión twb) donde se han realizado todas las tareas del guion de prácticas.

1.9. Instrucciones de envío

Envíe los Ficheros a través de Moodle **antes de la fecha límite**.

La entrega tardía está permitida, pero habrá una **penalización del 10% por cada 12 horas** hasta un máximo de 48 horas de demora.