Trabajo de Fin de Grado

Grado en Ingeniería Informática

Big Data y la Visualización en el ámbito Educativo

Big Data and visualization in Education

Yeray Pérez Peraza

La Laguna, 2 de Julio de 2017

D. **Dagoberto Castellanos Nieves**, con N.I.F. 79234766-L profesor Contratado Doctor adscrito al Departamento de Ingeniería Informática y de Sistemas, como tutor

C E R T I F I C A (N)

Que la presente memoria titulada:

"Big Data y la Visualización en el ámbito Educativo"

Ha sido realizada bajo su dirección por D. **Yeray Pérez Peraza**, con N.I.F. 54056611-X.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 2 de Julio de 2017.

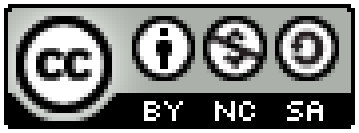
Agradecimientos

Dagoberto Castellanos Nieves

Luz Marina Moreno

Coromoto León

Licencia



Resumen

El objetivo de este trabajo ha sido el estudio del estado del Big Data el cual se encuentra en constante crecimiento, también se ha documentado el concepto de visualización ligado a Big Data y para terminar con los estudios de conceptos se ha tratado el tema de la educación.

Tras estudiar el concepto Big Data se ha procedido a buscar técnicas de visualización, con el fin de indicar la mejor técnica para visualizar los datos contenidos en nuestros data set.

Una vez estudiadas las técnicas de visualización podemos organizar y estructurar los datos contenidos en nuestros data sets. Aplicando las técnicas predictivas anteriormente estudiadas a los datos estructurados, para así obtener resultados.

Al disponer de las técnicas de visualización analizadas podemos realizar el preprocesamiento de los datos. Con el paso anterior realizado podemos nutrir a las herramientas de los datos estructurados para obtener una representación optima de los datos.

Palabras clave: *Big Data, visualización, educación, datos, técnicas de visualización.*

Abstract

The objective of this work has been the study of the real state of the discipline of Big Data, which is constantly growing. It has also been documented the concept of visualization connected to Big Data, and, to finish with studies of concepts, there has been talked about education.

After documenting the Big Data concept, we have proceeded to search for visualization techniques, in order to indicate the best one technique to visualize the data of the contents in our datasets.

Once studied visualization techniques can organize and structure the data in our data sets, we will apply predicative techniques and then represent them with the visualization techniques.

By having visualization techniques analyzed, it is already possible to perform the preprocessing of the data, with the previous step done we can nurture tools structured data, to get an optimal representation of the data.

Keywords: *Big Data, visualization, education, data, visualization techniques.*

Índice

Índice de figuras

Índice de tablas

Capítulo 1. Introductorio

1.1 Introducción

Debido al desarrollo tecnológico, que se ha vivido en la última época con el nacimiento de las redes sociales, los dispositivos móviles, los sensores, añadir internet a un amplio abanico de cosas y sobretodo la llegada de la tecnología a los ciudadanos, se ha generado un gran volumen de datos (Big Data), así nace la necesidad de poder analizar dichos datos para transformarlos en información [1].

Para tratar este volumen de datos el software y el hardware existente no cumplían con los requisitos, actualmente con el abaratamiento de los costes de producción es posible analizarlos. Ahora el problema reside en cómo analizar eficazmente los datos. También nace el inconveniente de como poder visualizar los datos de forma que sea sencillo e intuitivo.

Nuestro principal reto, es identificar metodologías existentes para analizar los datos y su visualización. El siguiente paso es seleccionarlas, eligiendo la más adecuada. Y también proponer una metodología que pudiera satisfacer las expectativas.

Con vistas a comparar o proponer metodologías necesitamos saber que técnicas de visualización son más adecuadas, para el fenómeno Big Data. A posteriori de realizar el análisis de dichas técnicas, se han detectado un grupo de técnicas que acorde a nuestros criterios pueden ser óptimas, puesto que pudieran hacer la información más accesible a cualquier usuario final, independientemente del nivel de conocimiento.

En consideración al reto principal, se debe documentar las diferentes técnicas predictivas que existen para aplicarlas a nuestros datos, por lo cual se ha de realizar una búsqueda de las principales técnicas y sus ámbitos de aplicación.

Con las técnicas de visualización ya definidas es hora de recorrer la

documentación para buscar distintas herramientas que nos permita crear una metodología, definiendo una serie de pasos para poder llegar a nuestro objetivo final.

El ámbito de estudio será el educativo, puesto que lo consideramos un ámbito interesante, no solo por el volumen del presupuesto de los distintos países, sino también puesto que la educación influye en el desarrollo de las personas, sus conocimientos, la cultura y la sociedad en general.

1.2 Introducción al Big Data

Debido al nacimiento de internet, sobre todo a la llegada de las redes sociales y con ello la generación de grandes volúmenes de datos, surge el concepto Big Data. El cual lleva asociado el estudio de los datos, visualización y representación de los mismos.

El fenómeno Big Data, el cual viene marcado por las 5 “v’s” las cuales definen las características de Big Data [2]. Nos apoyaremos en esta información para definir Big Data, con lo cual obtenemos una definición exacta y completa. En la siguiente ilustracioón podemos observar las caracteristicas que cumple el concepto Big Data.

En nuestra opinion la definición que mejor se adapta es la de 5 “v’s”, aunque tambien existen definiciones alternativas [3].

La primera es de las 3”v’s”, que simplemente tratan Volumen, Variedad y Velocidad, para nuestro criterio es una definición muy pobre de un fenomeno como el Big Data.

Otra definicion es de las caractersiticas Big Data es de las 4“v’s”, que añade la Veracidad a la definicion anterior, lo cual es importante tener conocimiento que los datos son reales.

La variedad de definiciones no solo acaba con 3 ,4 ó 5 “v’s”, existe la definición de las 7”v’s” para Big Data, la cual define las caracteristicas Big Data con Volumen , Velocidad, Variedad de los datos, Veracidad de los datos, Viabilidad , Visualización de los datos, Valor de los datos.

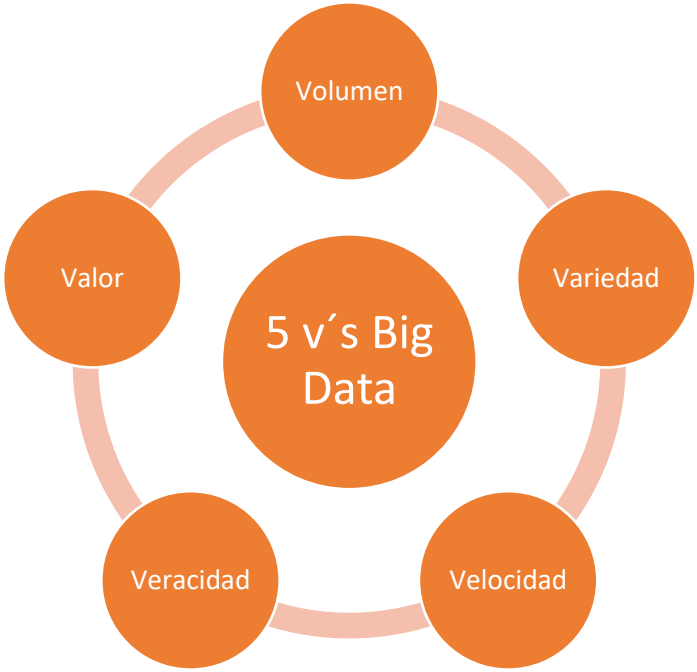


Ilustración 1: Big Data 5 v´s

Estas son las cualidades de la definición 5 "v's":

- **Volumen:** Se trabaja con grandes volúmenes de información, se llega al nivel de Peta bytes, Exabytes o Zettabytes.
- **Variedad:** Los datos con los que se trabaja son muy diversos y no tienen ningún tipo de estructura común.
- **Velocidad:** la información se genera a tal velocidad que es imposible gestionarla con bases de datos convencionales, la información se genera al segundo por lo que las personas y las empresas quieren actualizarla a dicha velocidad.
- **Veracidad:** Tener conocimiento de que los datos con los que estamos trabajando son reales y fieles a la realidad, es decir, tenemos la certeza de que los datos son reales o simplemente son todos.
- **Valor:** Los datos en sí tienen valor, aunque no solo monetario, pero sí tienen mucho valor informativo.

En [4] nos indica que podemos añadir otra "uve", la Visualización, puesto que la mayoría de los autores la excluye de Big Data, seguiremos su criterio.

Es importante la visualización puesto que conforma una parte importante de nuestro proyecto, pero la mayoría de los autores no incluye la visualización dentro de la definición de Big Data, seguiremos dicho criterio y trataremos la visualización como

un apartado aparte.

1.3 **Visualización**

Definiendo el concepto de visualización damos entrada al apartado sobre dicho tema.

Se entiende visualización como, representar mediante imágenes ópticas fenómenos de otro carácter; p. ej., el curso de la fiebre o los cambios de condiciones meteorológicas mediante gráficas, los cambios de corriente eléctrica o las oscilaciones sonoras con el oscilógrafo, etc. [5].

El objetivo a lograr con la visualización es empleando técnicas de procesamiento y visualización de datos simplificar la información mostrada, de tal forma que se pueda entender fácilmente.

Dichos resultados deben ser entendibles por los usuarios, independientemente de sus conocimientos, con la única ayuda de una imagen o un gráfico.

Es importante entender que la visualización de datos no es tarea únicamente de expertos en la materia tales como cartógrafos, programadores o estadísticos, lo que sí importa es saber que la visualización es tan relevante como los datos en sí además de las características ligada a la visualización tales como las escalas de los gráficos, las leyendas de los mapas, etcétera [4].

La información relevante de esta cita, radica en que es sumamente importante que la visualización sea factor clave a la hora realizar un análisis Big Data.

La mejor manera de entender la visualización es con el símil de que estamos buscando un tesoro o información oculta(a veces no sabemos que estamos buscando), debemos buscar las pistas como si se tratara de un crimen, hasta llegar a encontrar los datos realmente importantes para luego representarlos [4].

De esta afirmación obtenemos que no podemos predefinir los resultados a obtener debemos aplicar los pasos oportunos para llegar a un objetivo final, mayoritariamente incierto.

1.4 **Educación en el Big Data**

Entendemos como educación el proceso por el cual se trasmiten conocimientos a una persona para que esta adquiera una determinada formación [6].

En el ámbito educativo la correcta visualización de la información se convierte en fundamental, debido a que representando correctamente la información nos podemos garantizar que el proceso educativo se realice con calidad.

En lo referente a educación, vamos a hacer referencia al ámbito de educación como el espacio donde obtenemos los datos de nuestro data set para mejorar la relación entre enseñantes y alumnado. Asimismo, gracias al uso de los portales o plataformas educativas, se generan un gran volumen de datos.

Los datos referentes a educación con los que vamos a trabajar nos los proporciona la UNESCO, en [7] obtenemos los datos referentes al porcentaje de la población mayor de 25 y menor de 64 años que estudian a cualquier nivel, de todos los países miembros a dicha organización y [8] podemos leer porque es importante la educación para esta organización y los planes de futuro para escolarizar el mayor número posibles de personas.

Un problema derivado de la disciplina Big Data es la discriminación que pueden sufrir los alumnos, profesores o directivos que obvien los resultados o simplemente no realicen la metodología aquí dispuesta, es decir, si alguien no quiere, desea o decidan libremente no utilizar Big Data, serán discriminados de los que sí lo emplean por lo que desaprovechan los grandes beneficios de usar Big Data.

También hay que ser consciente de que no todos los profesores están capacitados tecnológicamente para emplear Big Data, por lo que sería ideal teorizar los para que sean capaces de aprovechar esta metodología.

Otro problema derivado de Big Data es la discriminación que pueden sufrir los alumnos al estudiar los valores anteriores a ellos, en otros términos, si por ejemplo existen antecedentes de que los alumnos de más de 25 años, son hombres y además estudien menos de cuatro horas semanales, van a suspender en un noventa y cinco por ciento de las veces, es injusto decir que los alumnos que cumplan estas características

no puedan cursar una asignatura porque tienen un 5 por ciento de posibilidades de aprobar.

1.5 Objetivos del proyecto

Los objetivos que persigue este proyecto se presentan a continuación:

1. Estudiar el estado del arte relacionado con Big Data y visualización.
2. Buscar, definir y localizar las diferentes técnicas de visualización y técnicas predictivas.
3. Buscar y comparar metodologías existentes, para visualizar los datos con técnicas predictivas de Big Data.
4. Proponer metodología acorde a las necesidades encontradas.
5. Realizar caso de uso.
6. Crear documentación técnica y la memoria del proyecto.

Capítulo 2. Big Data

Nos toca analizar y explicar las distintas características que componen el concepto Big Data, que definimos en el apartado anterior. En los apartados siguientes comprenderemos, los por menores de dicho concepto.

2.1 Retos del Big Data

Empezaremos por explicar los retos que conlleva implementar técnicas Big Data [9].

1. **Dar sentido a la gran cantidad de datos.**
   Es necesario tener las herramientas óptimas para poder dar sentido al gran volumen de datos que se generan por la bajada de costes en hardware y de las fuentes de datos.
2. **La comprensión de una variedad cada vez mayor de datos.**
   Es de vital importancia poder analizar los datos, ya sean relacionales o no. Es importante saber que un alto porcentaje de los datos no están estructurados (cerca del 85%).
3. **Habilitación de análisis en tiempo real de los datos.**
   Actualmente la mayor generación de datos (Twitter, Facebook, Instagram,...) generan volúmenes de datos sin parangón y todo esto en tiempo real, toda esta información no se puede analizar con efectividad con las técnicas normales de procesamiento de datos.
4. **Formar profesionales cualificados.**

   Tenemos la problemática que la sociedad actual no está preparada para afrontar los requisitos generados por la gestión que se debe realizar a los datos relacionamos con Big Data.

   Solucionar este reto consiste en formar profesionales especialistas en el análisis de los datos para que creen o usen herramientas de gestión de la información derivada de procesar los datos. En este entorno es donde las tecnologías de la información deben dar el salto. [10]

A modo de resumen podemos concluir que la tecnología asociada a Big Data y los objetivos de negocios deben coordinarse para ser capaces de obtener información de calidad.

2.2 Análisis de los datos

Los datos deben superar un análisis exhaustivo, para revelar la información que contienen los datos, para generar conocimiento.
El valor potencial de los datos no solo radica en disponer de ellos sino en saber organizarlos, refinarlos para convertirlos en información relevante, el valor que se le da a los datos después del análisis es para incrementar la capacidad de innovar y obtener ventaja sobre los competidores [10].

Gestionar correctamente los datos debe generar jurisprudencia en las empresas, la ciudanía y sobre todo en las administraciones ya sean públicas o privadas puesto que analizar correctamente los datos se debe convertir en un activo para la sociedad [10].

En [10], se proponen las fases del análisis (Ilustración 2)**.**

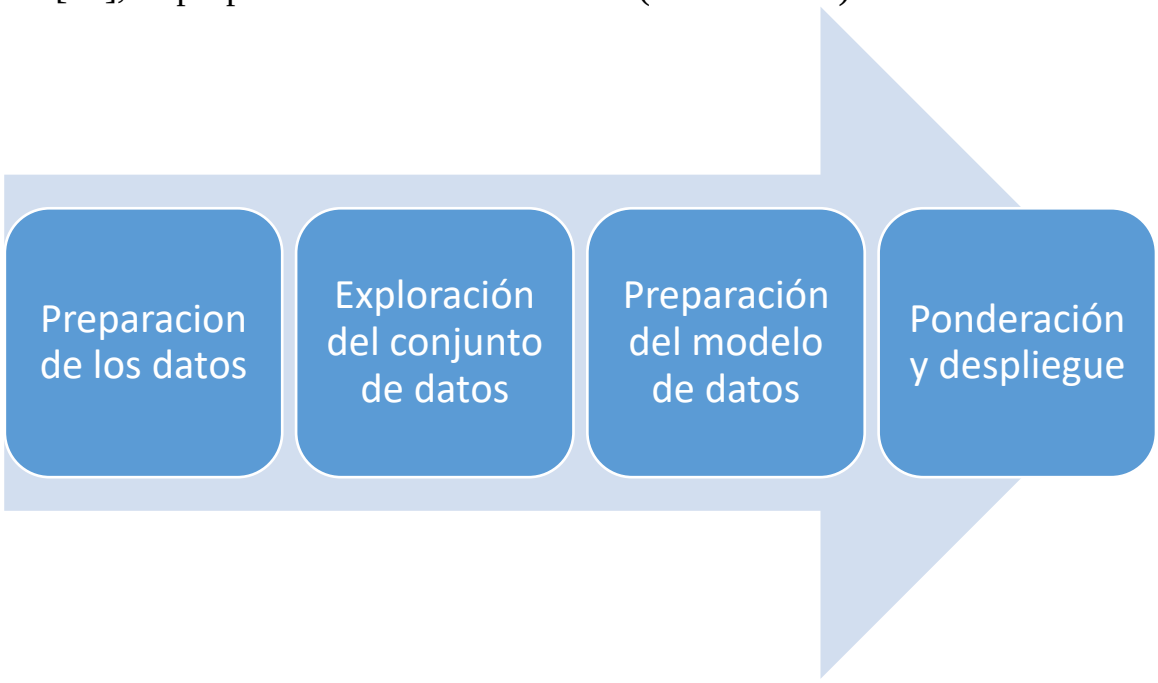


Ilustración 2: Fases de análisis

2.3 Tecnologías empleadas en el Big Data

Las tecnologías asociadas a Big Data son diversas tales como Hadoop, MapReduce,

MongoDB, Cassandra y el lenguaje R [11].

MapReduce es un modelo de programación utilizado por Google para dar soporte a la computación paralela sobre grandes colecciones de datos en grupos de computadoras y al commoditycomputing.

Apache Hadoop es un framework de software que permite a las aplicaciones trabajar con miles de nodos y petabytes de datos. Hadoop se inspiró en MapReduce.

MongoDBes un sistema de base de datos No SQL orientado a documentos, desarrollado bajo el concepto de código abierto.

Apache Cassandra es una base de datos No SQL distribuida y basada en un modelo de almacenamiento de «clave-valor».

Lenguaje R podemos programar técnicas predictivas orientadas a Big Data, usar paquetes que dispongan de dichas técnicas o incluso podemos manejar programas para procesamiento de Big Data.

Es obligado destacar que el lenguaje R, soporta una gran variedad de ámbitos de trabajo entre los cuales se encuentra Big Data.

2.4 **Importancia del Big Data**

En lo referente a la importancia lo mejor es hacer nuestra la frase que dice “la información es poder”, con esta reflexión debemos plantearnos que cualidad que hace importante a Big Data, es la cantidad cuanta más información más importante es Big Data. Debemos indicar que al definir las “5´v” de Big Data, una de ellas era el volumen, que habla concretamente de esta reflexión.

Pero la importancia del Big Data, no solo radica en el volumen, aunque es muy importante, también debemos saber que los datos sean veraces, es decir que sean reales o contrastados.

Ahora que sabemos que es importante en Big Data, podemos indicar que la importancia de la Big Data, radica en que es una buena fuente de información, donde podemos

extraer muchos datos del entorno. Big data no solo es importante para obtener información, sino que sabiendo cómo se comporta el pasado y el presente, con técnicas Big Data podemos predecir el futuro.

2.5 Propiedades sistema Big Data

Antes de terminar el capítulo es conveniente saber que propiedades que debe cumplir un sistema Big Data, con el objetivo de conocer que características tienen dichos sistemas [12].

1. **Robusto y tolerante a fallos.** Capacidad para un sistema para seguir trabajando, aunque falle algún elemento del sistema.
2. **Baja latencia de lectura y actualización.** Inexistencia de la necesidad de leer constantemente los datos o actualizarlos.
3. **Escalable.** Poder aumentar los componentes sin tener que reestructurar el sistema.
4. **General.** Que se pueda aplicar el sistema a un gran abanico de datos.
5. **Extensible.** Se pueda aplicar el sistema a diferentes ámbitos.
6. **Mínimo mantenimiento.** No requiera un alto gasto en mantenimiento de los elementos del sistema.
7. **Depurable.** Se pueda eliminar fácilmente los errores que puedan surgir.

Capítulo 3. Principales Técnicas de Visualización

La visualización correcta o efectiva de la información está condicionada por diversos elementos. Aunque, uno de los más relevantes es la correcta selección de las técnicas de visualización. En este trabajo hemos recopilado las técnicas más relevantes en este ámbito.

A continuación vamos a mostrar una tabla con las características de la visualización de cada técnica propuesta. [4] [13].

ID	Técnica	Comparing categories	Assessing hierarchies and part-to-whole relationships	Showing Changes over time	Plotting connections and Relationships	Multidimnsional Variables	Hierarchies	Maps	Add
0	Diagrama de barras	✓	×	×	×	✓	×	×	2
1	Grafico Circular	×	✓	×	×	✓	×	×	2
2	Gráfico de Líneas	×	×	✓	×	✓	×	×	2
3	Diagrama de dispersión	×	×	×	✓	×	×	×	1
4	Mapa coroplético.	×	×	×	×	✓	×	✓	2
5	Gráfico de Puntos	✓	×	×	×	✓	×	×	2
6	Diagrama de barras flotante	✓	×	×	×	✓	×	×	2
7	Histograma	✓	×	×	×	✓	×	×	2
8	Gráfico de glifos	✓	✓	×	×	✓	×	✓	4
9	Gráfico de área grande	✓	×	×	×	✓	×	×	2
10	Pequeños múltiples	✓	✓	×	×	✓	×	×	3
11	Nubes de palabras	✓	×	×	×	×	✓	✓	3

12	Gráfico de barras apilado	✓	✓	×	×	✓	×	×	3
13	Gráfico de waffles	×	✓	×	×	✓	×	×	2
14	Mapa de árbol	×	✓	×	✓	✓	×	✓	4
15	Diagrama de embalaje en círculo	×	✓	×	×	×	×	×	1
16	Jerarquía de burbujas	×	✓	×	✓	✓	×	✓	4
17	Jerarquía de árbol	×	✓	×	✓	✓	×	✓	4
18	Chispas	×	×	✓	×	×	×	×	1
19	Gráfico de área	×	×	✓	×	×	×	×	1
20	Gráfico de flujo	×	×	✓	×	×	×	×	1
21	Gráfico de velas	×	×	✓	×	✓	×	×	2
22	Gráfico de código de Barras	×	×	✓	×	✓	×	×	2
23	Mapa de flujo	×	×	✓	×	✓	×	×	2
24	Diagrama de Burbujas	×	✓	×	✓	✓	×	✓	4
25	Matriz de diagrama de dispersión	×	×	×	✓	✓	×	✓	3
26	Gráfico de la matriz	×	×	×	✓	✓	×	✓	3
27	Diagrama de acordes	×	×	×	✓	✓	×	✓	3
28	Diagrama de Red	×	×	×	✓	✓	×	✓	3
29	Mapas de puntos	×	×	×	×	✓	×	✓	2
30	Mapa de puntos (Burbujas)	✓	×	×	×	✓	×	✓	3
31	Mapa topológico	✓	×	×	×	✓	×	✓	3
32	Cartograma	✓	×	×	×	✓	×	✓	3
33	Cartograma Dorling	✓	×	×	×	✓	×	✓	3
34	Mapa de conexiones de red	✓	×	✓	×	✓	×	✓	4

35	Grafos	×	×	×	×	×	✓	×	1
36	Redes de Grafos	×	×	×	×	×	✓	×	1
37	Diagrama de Flujos	×	×	×	×	✓	✓	×	2
38	Mapa mental	×	×	×	×	×	✓	×	1
39	Diagrama de Procesos	×	×	×	×	×	✓	×	1
40	Diagrama analítico	×	×	×	×	×	✓	✓	2
41	Organigrama	×	×	×	×	×	✓	✓	2

Tabla 1: Características de las técnicas de visualización

En la siguiente tabla mostramos las técnicas mejor valoradas para Big Data del listado de las técnicas anteriores.

Representación recomendada para Big Data									
ID	Técnica	Comparing categories	Assessing hierarchies and part-to-whole relationships	Showing Changes over time	Plotting connections and Relationships	Multidimnsional Variables	Hierarchies	Maps	Add
0,6, 7,9	Visualización de volúmenes	✓	×	×	×	✓	×	×	2
20	Diagramas de Flujo	×	×	×	×	✓	✓	×	2
35	Grafos	×	×	×	×	×	✓	×	1
42	Pixel Bars	✓	×	×	×	×	✓	×	2
43	Infografía	✓	✓	×	✓	✓	✓	✓	6
11	Nubes de palabras	×	×	×	×	×	✓	✓	2
14	Mapa de árbol	×	✓	×	✓	✓	×	✓	4
16	Jerarquía de burbujas	×	✓	×	✓	✓	×	✓	4
28	Diagrama de Red	×	×	×	✓	✓	×	✓	3
31	Mapa topológico	✓	×	×	×	✓	×	✓	3
32	Cartograma	✓	×	×	×	✓	×	✓	3
33	Cartograma Dorling	✓	×	×	×	✓	×	✓	3

Tabla 2: Características de las técnicas de visualización especial para Big Data

Leyenda de las Tablas:

Cumple	✓
No cumple	×

Puesto la gran complejidad de representar los resultados de los estudios sobre Big Data, se han especificado varias técnicas que engloban las mejores opciones de presentación las cuales van desde las más sencillas como los diagramas de barras, pasando por las representaciones en forma de árbol como los grafos hasta llegar a los infografías, las cuales son las técnicas mejor valoradas para representar los datos puesto que engloban muchas técnicas de presentación por lo cual indican de manera clara los resultados [14] [15].

Tras realizar la suma ponderada de los técnicas que forman la tabla hemos realizado un diagrama con las valoraciones más altas, para ver que tecnica puede ser la más adecuada para añadirla a nuestra metodología de trabajo.

Id	Técnica	Total
44	Infografía	6
8	Gráfico de glifos	4
14	Mapa de árbol	4
17	Jerarquía de árbol	4
24	Diagrama de Burbujas	4
34	Mapa de conexiones de red	4

Tabla 3: Técnicas de visualización mejor valoradas

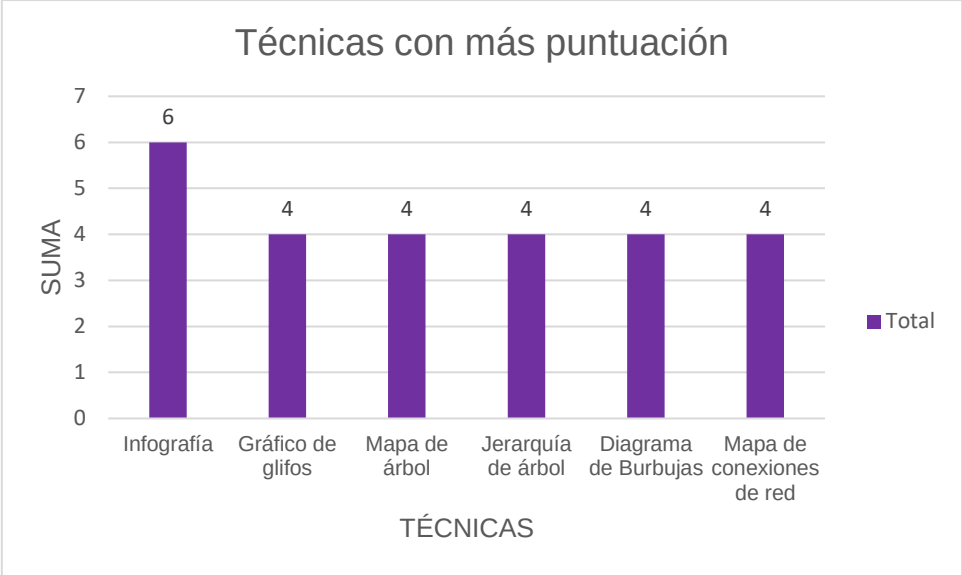


Ilustración 3: Comparativa de las técnicas con más características

De las técnicas con mayor valoración destaca las infografías las cuales agrupan en una sola imagen varias formas de visualización, es decir, en una infografía podemos ver varios gráficos, algún mapa o alguna jerarquía de árbol. Las demás técnicas con mayor valoración observamos un empate a cinco entre formas de visualización con valor de cuatro, podemos agrupar tres de ellas por similitud, estas serían mapa de árbol, jerarquía de árbol y mapas de conexiones de red, luego nos quedan dos técnicas con la misma valoración una son los diagramas de burbujas y la otra los gráficos de glifos. A continuación, vemos un ejemplo de cada una de las técnicas a las que nos referimos:

Ilustración 4: Técnicas de visualización agrupadas

También nos interesa saber que característica está más presente dentro de las elegidas para comparar que técnicas es la más adecuada, como podemos observar en la Ilustración 5: Valores de las características, la característica más repetida es la de "Comparing categories", seguida de "Multidimensional Variables", por lo que podemos determinar que existe un alto número que pueden comparar categorías con varias variables.

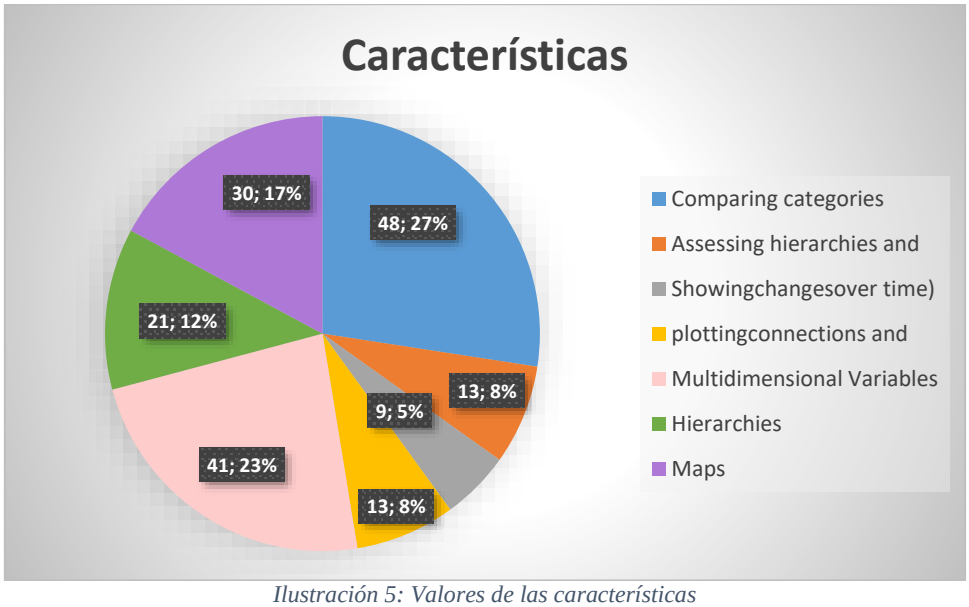


Ilustración 5: Valores de las características

En la siguiente infografia se explica que componente debe tener una infografia, nos explican que debe tener una fase de diseño y nos indica varios ejemplos de las posibles tecnicas de visualización a emplear [16].

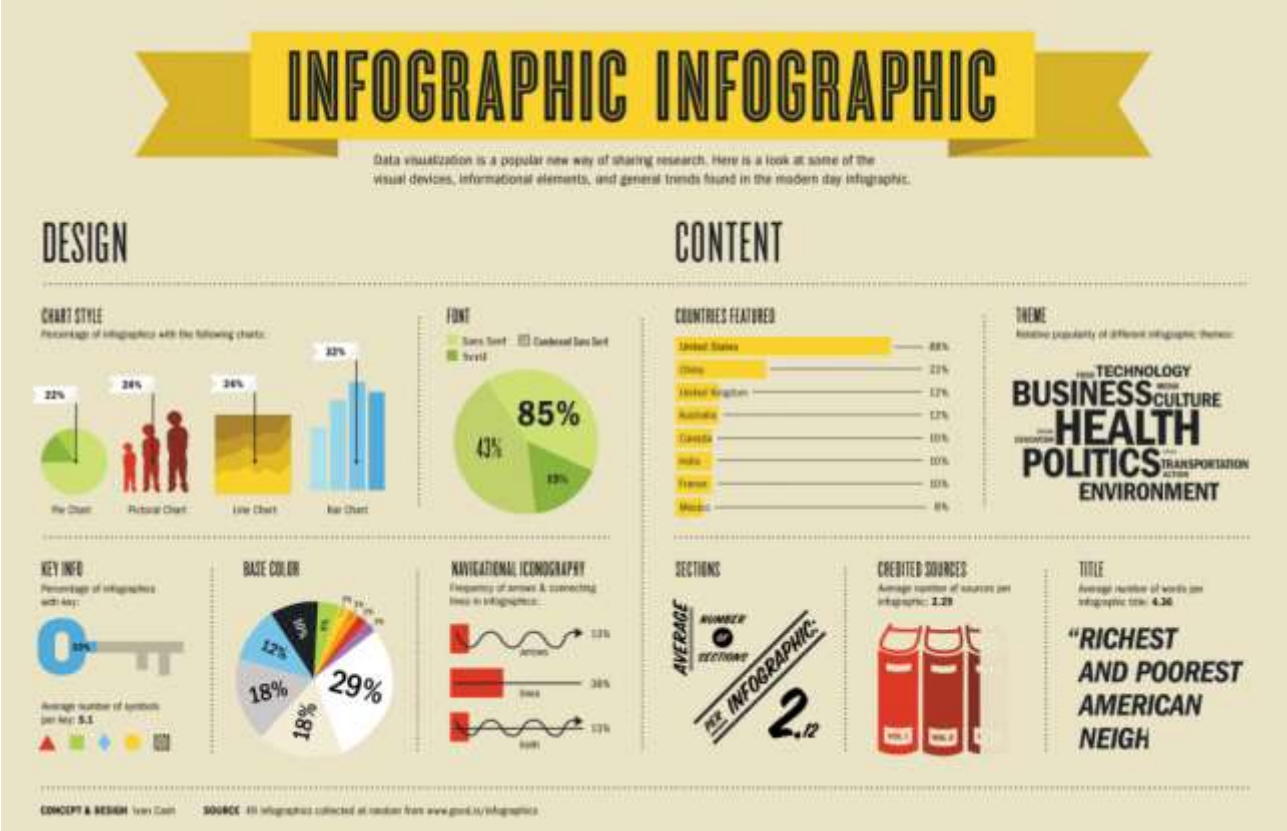


Ilustración 6: Infografía de ejemplo

Capítulo 4.
Metodología propuesta

Entendamos una metodología como un conjunto de paso, métodos o indicaciones que debemos seguir para conseguir un objetivo final [17].

4.1 Metodologías existentes

En [18]. , podemos ver las diferentes metodologías que podemos aplicar al ámbito educativo, el mismo que trata nuestro proyecto, seguidamente se las mostraremos.

4.1.1 Técnicas estadísticas predictivas.

Las técnicas estadísticas predictivas tienen como objetivo predecir el comportamiento de un aspecto de los datos (variable dependiente) como combinación del resto de características (variables independientes).

Una de las técnicas predictivas más usadas es el teorema de Bayes, a continuación les mostramos una metodología para realizarlo [19] [20]:

1. Se percibe y se evalúa una situación a la luz de las evidencias y acontecimientos observados.
2. Se formulan los escenarios probables / hipótesis alternativas y se le asignan unas probabilidades subjetivas iniciales. Tales escenarios deben cumplir con la condición de exahustividad y exclusión mutua.
3. Se inicia el proceso de seguimiento y monitoreo de todos los eventos (acontecimientos), hechos que inciden en el direccionamiento de las tendencias.
4. Con base en el registro de eventos (evidencias) se ajustan por el método de bayes las probabilidades de ocurrencia asignadas a cada escenario.
5. Una vez hecho los cálculos tomando como base los juicios de valor de los analistas y expertos se hacen los gráficos de tendencias.
6. Visualizando los gráficos de tendencias en cuanto a las posibilidades de

ocurrencia de cada escenario, se evalúa la necesidad de dar “la alerta”.

7. De ser requerido dar “el alerta”; la misma tendrá que fundamentarse de manera lógica y convincente en las evidencias obtenidas hasta el momento. Tal “alerta” deberá servir de base para una toma de decisiones oportuna ante la situación planteada.

4.1.2 Descubrimiento de estructuras.

El objetivo en este caso es el descubrimiento de estructuras y patrones en los datos capturados. Abarca los diversos algoritmos de clustering y el análisis factorial.

Existen varias técnicas de clustering, a continuación les expondremos la técnica clustering jerárquico aglomerativo [21] [22]:

1. Creamos una matriz de similitud que contiene las distancias entre los distintos elementos a agrupar. En nuestro caso esta matriz se calcula a partir de la matriz de correlaciones.
2. Consideramos todas las agrupaciones posibles y elegimos la mejor según la matriz de similitud.
3. Se recalcula la matriz de similitud teniendo en cuenta el nuevo clúster formado. La distancia al nuevo clúster se calcula como la media de las distancias a los elementos que lo forman.
4. Volvemos al punto dos hasta alcanzar el resultado esperado.

4.1.3 Minería de relaciones.

La minería de relaciones trata sobre el descubrimiento de las relaciones entre las variables dentro de un conjunto extenso de datos. Inmediatamente de bajo les mostramos una metodología para ejecutar la minería de relaciones [23] [24]:

1. Obtener los datos.
2. Limpiar los datos.
3. Bodega de datos (almacén para los datos)
4. Selección de los datos.
5. Definir el objetivo del análisis.
6. Correlación de la información.

7. Evaluación de patrones.
8. Obtener el conocimiento de los datos.

4.1.4 Procesamiento del lenguaje natural.

El procesamiento del lenguaje natural es un conjunto de técnicas algorítmicas para analizar el lenguaje humano.

Con fin de proporcionar una metodología dedicada a este ámbito, les mostramos los siguientes métodos para llevarla a cabo [25] [26].

1. Recuperación de la información.
    a. Eliminación de coincidencia de las formas morfológicas de palabras.
    b. Eliminación de coincidencia de los sinónimos.
    c. Eliminación de palabras sin valor (No dan información).
2. Reconocimiento y la clasificación de entidades nombradas.
3. Mostrar los resultados.

Finalmente, con la información mostrada anteriormente, contamos con 5 posibles puntos de partida para obtener una buena metodología para el ámbito educativo.

4.1.5 Desarrollo por partes

Ahora entramos en metodologías agiles relacionadas con Big Data, en este caso el desarrollo por partes.

En [27], se propone realizar el desarrollo por partes:

1. Se realizan las ETL (extraer, transformar y cargar) pero sólo las de esa área, no todas.
2. Se crean los cubos OLAP (procesamiento analítico en línea) relativos a esa área.
3. Se crean los informes relativos a esa área.
4. Se crean los cuadros de mando relativos a esa área.
5. Validar con el usuario.
6. Podemos aplicar las correcciones al área desarrollada e implementar las nuevas

áreas con las nuevas especificaciones.
7. Obtenemos resultados a corto plazo, todo el mundo está implicado y nadie se olvida de que estamos realizando un proyecto y el objeto del mismo.
8. Las validaciones empiezan desde un primer momento y se hacen progresivamente. No descubrimos los errores cuando ya tenemos todo el sistema construido.

4.1.6 Ciencia de los datos prácticos

Ciencia de datos es el conjunto de prácticas sobre almacenamiento, gestión y análisis de conjuntos de datos lo suficientemente grandes que requieren de computación distribuida y los recursos de almacenamiento [28].

Otro ejemplo de metodologías lo podemos encontrar en [29], esta metodología está basada en la ciencia de datos:

1. Datos
    a. Carga de datos
    b. Exploración de datos
    c. Gestión de datos
2. Modelización
    a. Selección y evaluación de modelos
    b. Métodos de memorización
    c. Regresión lineal y logística
    d. Métodos no supervisados
3. Resultados
    a. Documentación e implementación
    b. Producción de presentaciones efectivas

4.2 Metodología propuesta

Por otra parte, después de documentar las metodologías existentes, nos hemos dado cuenta que ninguna de las documentadas anteriormente, son adecuadas a nuestras necesidades, por lo que es necesario definir nuestra propia metodología.

1. **Selección del conjunto de datos.**

   Debemos buscar un data set que contemple el mayor número de datos y variables del ámbito que vamos a estudiar, es imprescindible recordar que la mayoría de los datos no siguen una estructura fija por lo que tener un data set completo es muy importante para obtener unos resultados reales y eficientes.

2. **Análisis de las propiedades de los datos.**

   Con los datos o variables que disponemos en nuestra data set tendremos que valorar que propiedades de ellos son las realmente importantes para realizar el análisis.

3. **Preprocesamiento de los datos.**

   Antes de poder analizar los datos con los modelos de Big Data, debemos estructurarlos, unirlos o simplemente ordenarlos para poder dar consistencia a la información que va a recibir las técnicas de visualización y obtener resultados fiables.

4. **Seleccionar y aplicar las técnicas de visualización.**

   Dentro de la gran variedad de técnicas de visualización que hemos analizado previamente, tomaremos las mejor valoradas para el ámbito Big Data y de estas el usuario elegirá la que desee.

5. **Extracción de conocimiento.**

   Luego de realizar la técnica o técnicas predictivas elegidas, obtenemos información o conocimiento útil de los datos.

6. **Interpretación y evaluación.**

   Una vez extraída la información podemos dar significado a los datos representados, para finalmente evaluar si la técnica usada es la más adecuada para representar correctamente los datos.

Capítulo 5. Caso de Estudio

La realización del caso de uso radica en validar las técnicas de visualización con los distintos data set que vamos a explicar a continuación.

Para realizar el caso de uso emplearemos dos data set, proporcionados por la UNESCO [8] [7], el primero con datos tipo numéricos y el segundo es un corpus lingüístico.

En el primero [7], los datos representan la población de todos los países asociados a la Unesco, en intervalos de cinco años, desde 1870 hasta el 2010, para identificar las naciones se clasifican en las siguientes regiones:

- Economías avanzadas. (Advanced Economies.)
- Asia y el Pacífico. (Asia and The Pacific.)
- Europa del Este. (Eastern Europe.)
- América Latina y el Caribe. (Latin America and The Caribbean.)
- Oriente Medio y África del Norte. (Middle East and North Africa.)
- Africa Sub-sahariana. (Sub-aharan Africa.)

Los datos que la componen son los siguientes:

- **País (Country):** Indica el país al que pertenecen los datos.
- **Año (Year):** Indica el año desde 1870 hasta el 2010, en el que se tomaron los datos.
- **Grupo de edad (Age Group):** Está variable almacena el rango de edad de las personas estudiadas, abarca desde los 25 hasta los 64 años.
- **Sin escolarizar (No Schooling):** Porcentaje de la población en el rango de edad sin escolarizar.
- **Mayor nivel alcanzado (Highest level attained):** Almacena los porcentajes de la población representada, que ha alcanzado un cierto nivel educativo, que engloba los niveles primario, secundario y terciario.
- **Promedio anual de escolaridad total (Avg. Years of Total Schooling):** Porcentaje de la población, escolarizada ese año.
- **Promedio anual escolarizado en primaria (Avg. Years of Primary Schooling):** Porcentaje de la población, escolarizada en primaria ese año.
- **Promedio anual escolarizado en secundaria (Avg. Years of Secondary**

Schooling:) Porcentaje de la población, escolarizada en secundaria ese año.

- **Promedio anual escolarizado en terciaria (Avg. Years of Tertiary):** Porcentaje de la población, escolarizada en la universidad o estudios superiores ese año.
- **Población (Population):** Población desde 25 hasta 64 del año de los datos.
- **Región (Region):** Indica la región a la que pertenece el país estudiado.

	Region	Age	Avg.YearsofTotalSchooling	Avg.YearsofPrimarySchooling	Avg.YearsofSecondarySchooling	Avg.YearsofTertiarySchooling
1	Advanced Economies	1870	0.3987	0.3819	0.0155	0.0011
2	Advanced Economies	1875	0.4185	0.3992	0.0176	0.0014
3	Advanced Economies	1880	0.4464	0.4243	0.0204	0.0025
4	Advanced Economies	1885	0.4783	0.4532	0.0231	0.0036
5	Advanced Economies	1890	0.5186	0.4876	0.0291	0.0061
6	Advanced Economies	1895	0.5580	0.5212	0.0340	0.0102
7	Advanced Economies	1900	0.6098	0.5643	0.0418	0.0154
8	Advanced Economies	1905	0.6643	0.6095	0.0502	0.0216
9	Advanced Economies	1910	0.7205	0.6549	0.0594	0.0292
10	Advanced Economies	1915	0.7860	0.6999	0.0776	0.0385
11	Advanced Economies	1920	0.8527	0.7480	0.0947	0.0596
12	Advanced Economies	1925	0.9255	0.7960	0.1168	0.1008
13	Advanced Economies	1930	0.9984	0.8425	0.1404	0.1415
14	Advanced Economies	1935	1.0768	0.8899	0.1681	0.1857
15	Advanced Economies	1940	1.1545	0.9368	0.1959	0.2488
16	Advanced Economies	1945	1.2353	0.9819	0.2291	0.0000
17	Advanced Economies	1950	1.3281	1.0414	0.2578	0.0000
18	Advanced Economies	1955	1.3874	1.0714	0.2824	0.0013
19	Advanced Economies	1960	1.4577	1.0939	0.3249	0.0018
20	Advanced Economies	1965	1.5506	1.1206	0.3850	0.0026
21	Advanced Economies	1970	1.6730	1.1485	0.4650	0.0045
22	Advanced Economies	1975	1.8122	1.1576	0.5745	0.0074
23	Advanced Economies	1980	1.9565	1.1893	0.6680	0.0114
24	Advanced Economies	1985	2.0568	1.2064	0.7312	0.0165
25	Advanced Economies	1990	2.1954	1.2317	0.8222	0.0218
26	Advanced Economies	1995	2.3250	1.2533	0.9099	0.0296
27	Advanced Economies	2000	2.4425	1.2652	0.9912	0.0403

Ilustración 7: Fragmento del primer data set

El segundo data set también de la UNESCO [8] , donde se explican los planes de actuación de dicha asociación para replantear la educación para buscar un bien común. Esta data set es de tipo texto por lo que no disponemos de variables sino de un texto que trata de un tema específico, procesaremos los datos para obtener una nube de palabras más repetidas.

Para el caso de uso vamos a seguir la metodología dispuesta anteriormente, en la cual explicamos los paso a seguir.

Primero seleccionamos el conjunto de datos, vamos a trabajar con los distintos data set, en los cuales hemos asociado los países por la región a las que pertenezcan, con esta operación pre-procesamiento de los datos conseguimos estructurar los datos.

Segundo: Análisis de las propiedades de los datos, realizamos el estudio previo de las variables que consideramos más importantes, en nuestro caso son las siguientes:

- Región (Region).
- Año (Age).
- Promedio anual de escolaridad total (Avg. Years of Total Schooling)
- Promedio anual escolarizado en primaria (Avg. Years of Primary Schooling)
- Promedio anual escolarizado en secundaria (Avg. Years of Secondary Schooling)
- Promedio anual escolarizado en terciaria (Avg. Years of Tertiary Schooling)

Se han seleccionado estas variables, puesto que sus características las destacan sobre las demás, las variables "Region" y "Age", se seleccionan por identificar el año del estudio y la región a los que pertenecen los datos.
También se ha elegido las variables que informan sobre los porcentajes de escolarización totales y en los distintos niveles educativos.

Antes de empezar con los modelos predictivos, debemos pre-procesar los datos, el cual consiste en realizar un sumatorio de las variables según el año y la región a las que pertenecen los datos.

Tras estos pasos previos podemos empezar con el modelo predictivo.

El primer paso, radica en incorporar los datos al entorno estadístico (Lenguaje R, con el IDE Rstudio). Añadimos los datos de cada región por separado.

	Region	Age	Population..1000s.	Avg..Years.of.Total.Schooling	Avg..Years.of.Primary.Schooling	Avg..Years.of.Secondary.Schooling	Avg..Years.of.Tertiary..Schooling
3	Asia and the Pacific	1880	261342	0.0092	0.0080	0.0007	0.0000
4	Asia and the Pacific	1885	282535	0.0107	0.0095	0.0008	0.0000
5	Asia and the Pacific	1890	286708	0.0125	0.0115	0.0009	0.0000
6	Asia and the Pacific	1895	300412	0.0159	0.0141	0.0012	0.0000
7	Asia and the Pacific	1900	309440	0.0200	0.0180	0.0016	0.0000
8	Asia and the Pacific	1905	323891	0.0289	0.0262	0.0025	0.0001
9	Asia and the Pacific	1910	337568	0.0408	0.0369	0.0039	0.0001
10	Asia and the Pacific	1915	351561	0.0589	0.0517	0.0068	0.0006
11	Asia and the Pacific	1920	365794	0.0781	0.0670	0.0098	0.0012
12	Asia and the Pacific	1925	376600	0.1079	0.0902	0.0156	0.0020
13	Asia and the Pacific	1930	387653	0.1395	0.1142	0.0226	0.0027
14	Asia and the Pacific	1935	411087	0.1686	0.1354	0.0291	0.0035
15	Asia and the Pacific	1940	434083	0.1973	0.1571	0.0357	0.0044
16	Asia and the Pacific	1945	456011	0.2254	0.1790	0.0417	0.0050
17	Asia and the Pacific	1950	447251	0.2630	0.2088	0.0481	0.0060
18	Asia and the Pacific	1955	477638	0.3061	0.2430	0.0559	0.0073
19	Asia and the Pacific	1960	519732	0.3240	0.2521	0.0653	0.0082
20	Asia and the Pacific	1965	568666	0.3727	0.2884	0.0748	0.0098
21	Asia and the Pacific	1970	624009	0.4385	0.3312	0.0981	0.0115
22	Asia and the Pacific	1975	696292	0.5013	0.3703	0.1174	0.0138
23	Asia and the Pacific	1980	800865	0.5977	0.4309	0.1474	0.0192
24	Asia and the Pacific	1985	909134	0.6954	0.4817	0.1897	0.0240
25	Asia and the Pacific	1990	1041627	0.7895	0.5253	0.2323	0.0319
26	Asia and the Pacific	1995	1196743	0.8875	0.5718	0.2758	0.0400
27	Asia and the Pacific	2000	1343954	0.9651	0.6043	0.3111	0.0496
28	Asia and the Pacific	2005	1464361	1.0569	0.6364	0.3566	0.0637
29	Asia and the Pacific	2010	1592699	1.1346	0.6639	0.3999	0.0711

Ilustración 8: Datos de Asia y el Pacifico

El siguiente paso de la metodología consiste en validar el modelo predictivo, para ello aplicamos varios y comprobaremos el resultado obtenido.

Empezamos empleado la técnica para crear un mapa de calor donde medimos y comparamos los valores de los porcentajes de población que realizan estudios, con edades comprendidas entre 25 y 64 años.

Hemos realizado un mapa de calor por cada una de las regiones que forman el conjunto de los datos. A continuación, les mostramos dos ejemplos de dichos mapas de calor.

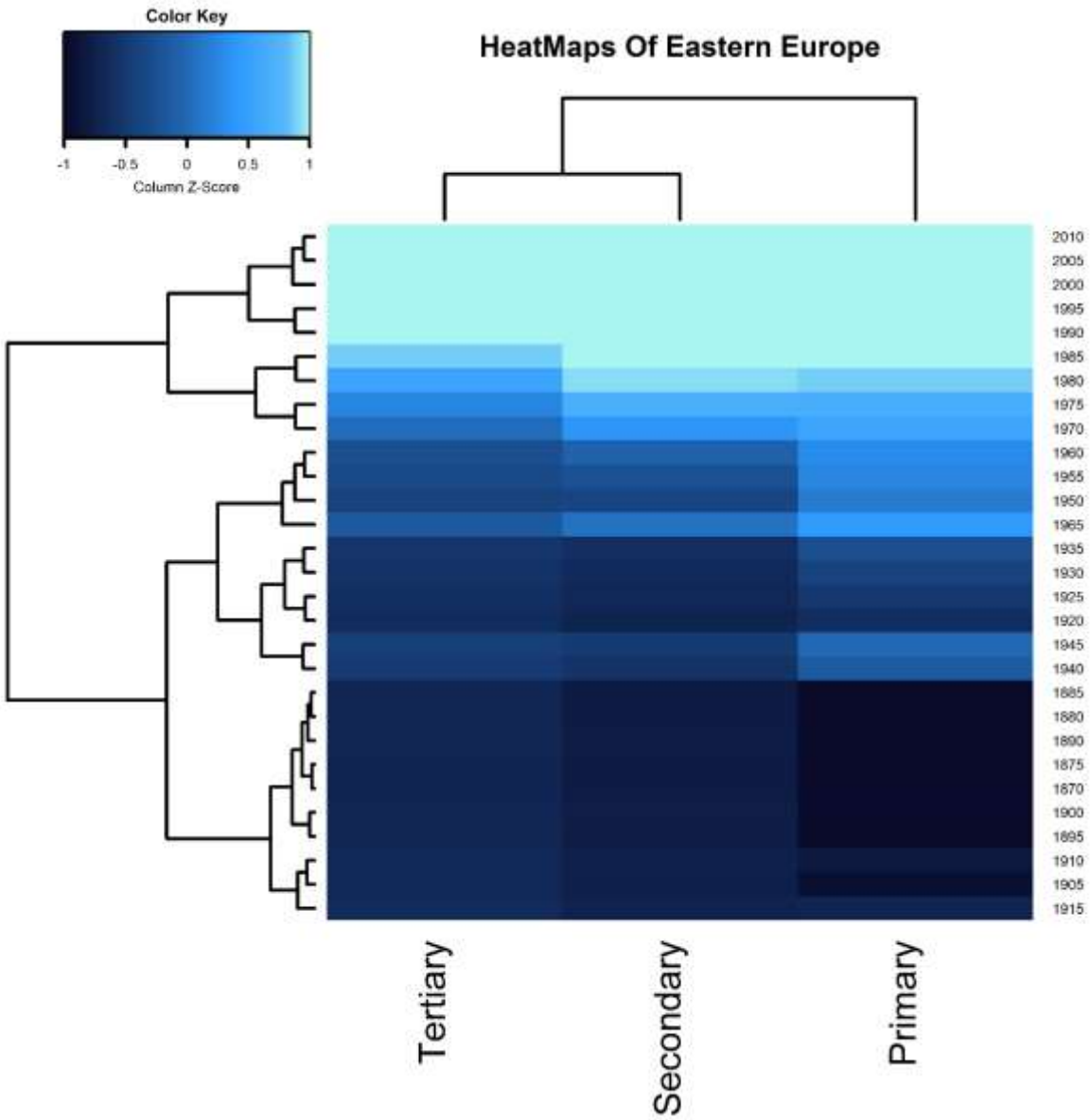


Ilustración 9: Mapa de calor de la región Europa del Este

Este grafico indica según la gama cromática que porcentaje de personas adultas que estudian en primaria, secundaria y terciaria en Europa del Este.

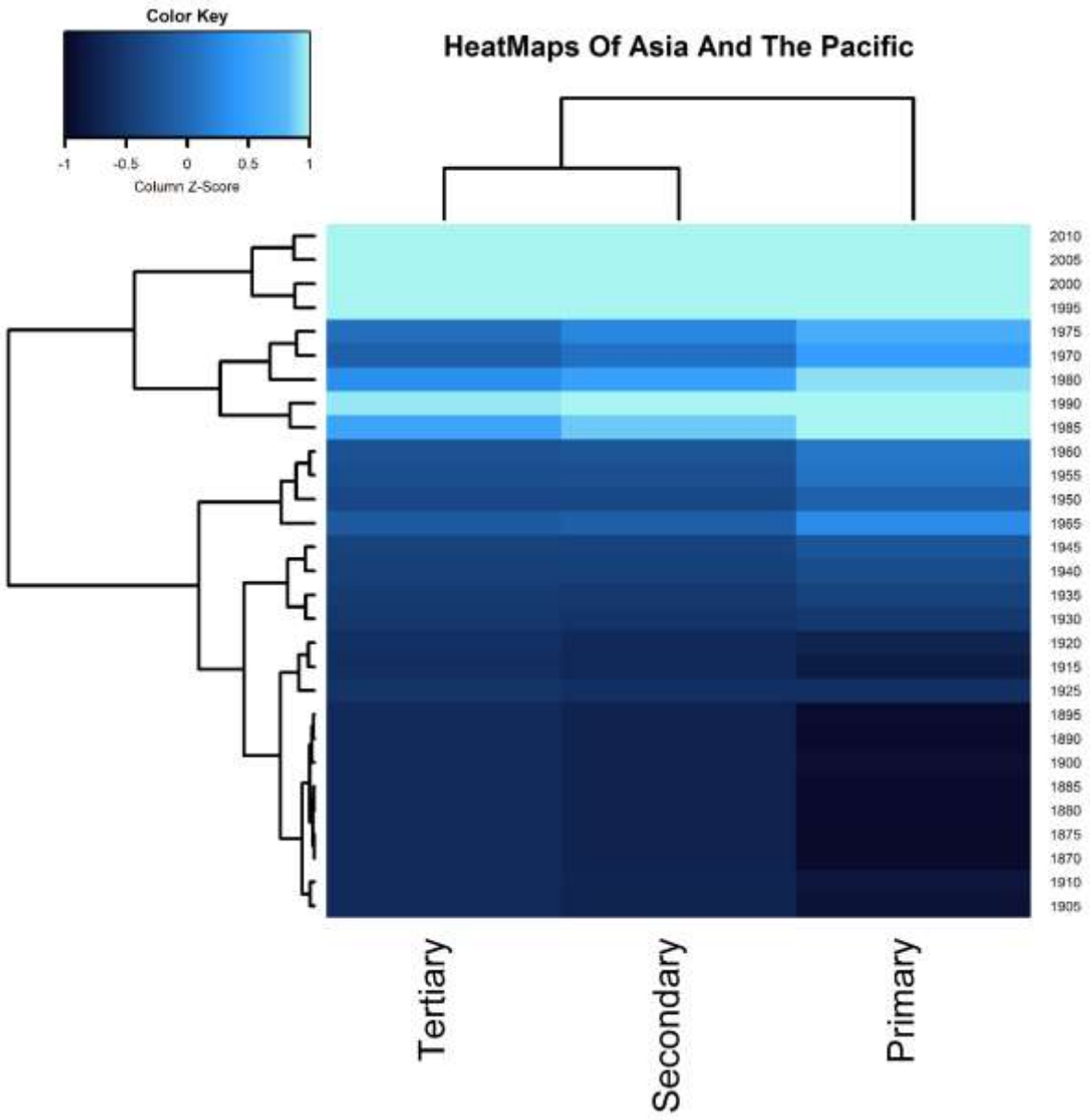


Ilustración 10: Mapa de calor de la región Asia y el Pacífico

Este grafico indica según la gama cromática que porcentaje de personas adultas que estudian en primaria, secundaria y terciaria en Asia y el Pacífico

Finalmente, con la ayuda del grafico generado podemos concluir que en los primeros años estudiados en todas las regiones el porcentaje de estudiados es bajo, en la actualidad esta tendencia continúa siendo la tónica predominante en las regiones subdesarrolladas.

Probaremos con otra técnica la creación de diagramas de puntos, formaremos un

ranking con el cual ordenaremos la información por los valores de las variables estudiadas. Con el fin de evaluar el resultado de aplicar este modelo creamos diagramas de puntos para todas las regiones del conjunto de los datos, para mejor su visibilidad en este documento mostraremos dos ejemplos de los datos.

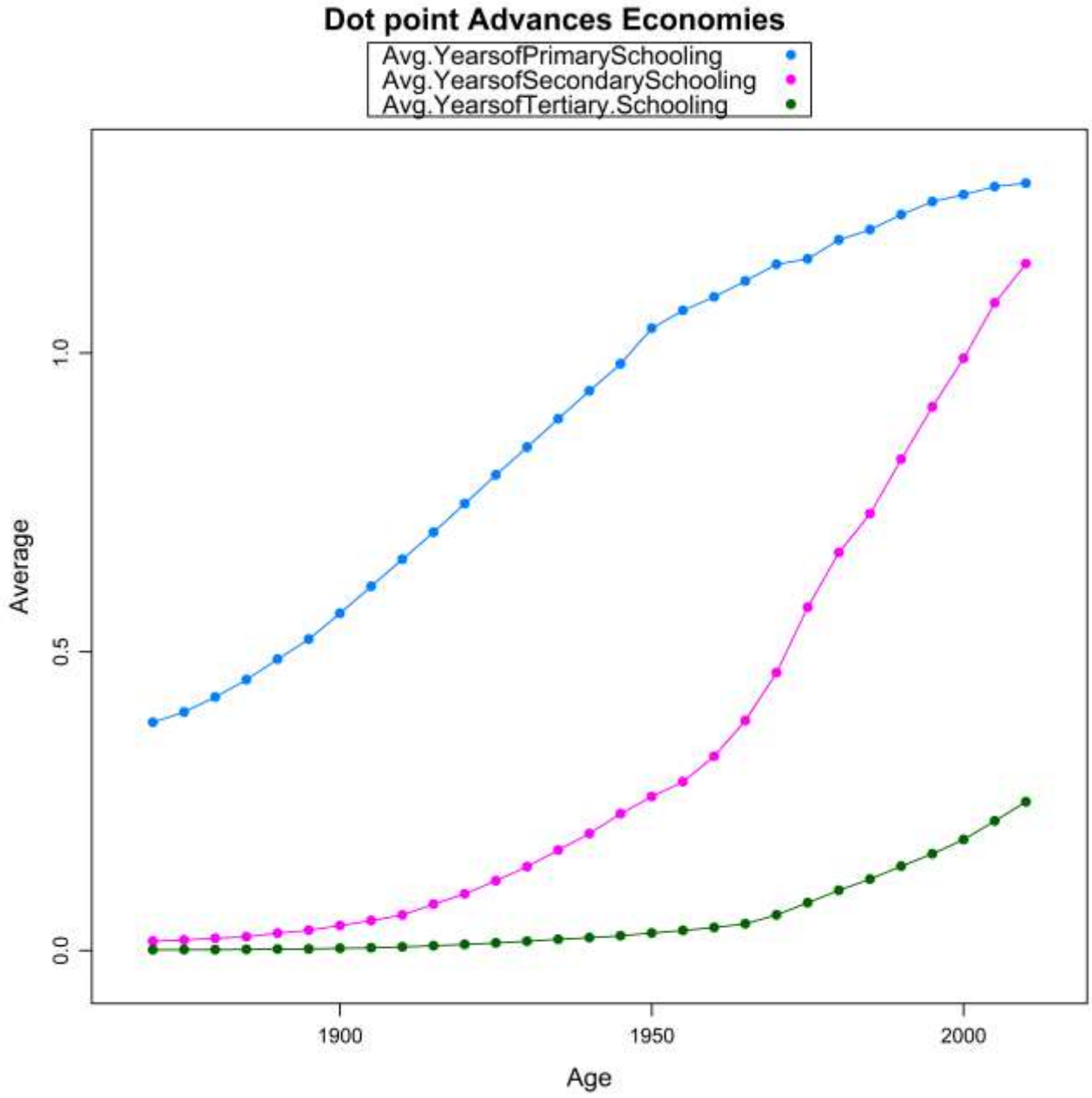


Ilustración 11: Diagrama de puntos de la región Economías Avanzadas

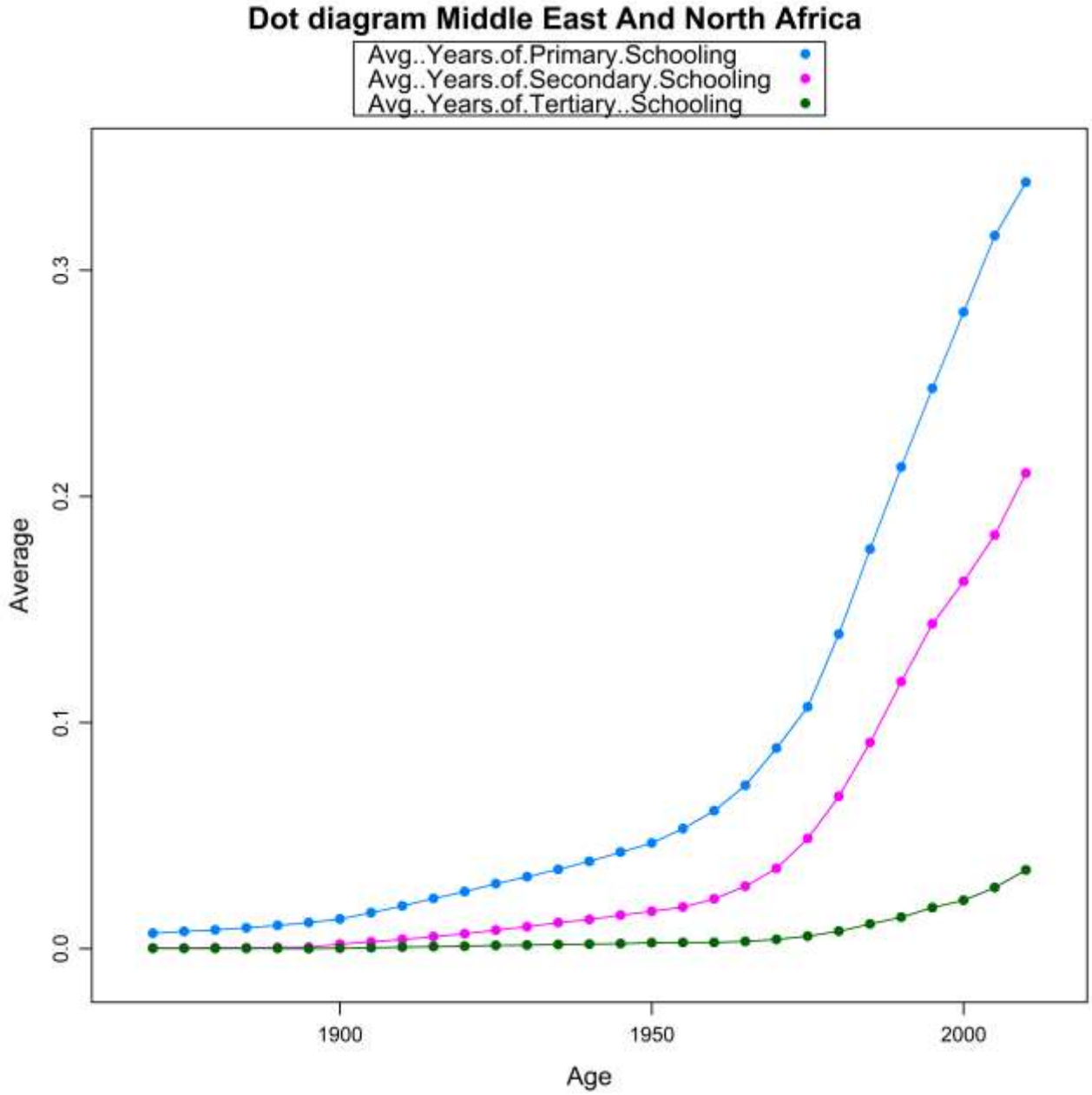


Ilustración 12: Diagrama de puntos de Oriente Medio y África del Norte

En los dos gráficos anteriores observamos con ayuda de las evoluciones que han seguido las variables con el paso de los años.

Como conclusión de realizar esta metodología podemos observar que con el paso del tiempo la población de 25 años a 64 años en el mundo que realiza estudios aumenta sobre todo en niveles primarios y secundarios, es decir, podemos concluir que las personas que estudian fuera de una edad regular aumentan con el paso de los años.

A continuación, aplicaremos otra técnica de visualización en este caso será visualización por volúmenes, donde podemos encontrar entre otros diagramas de barras, diagramas de barras flotantes o diagrama de pixeles.

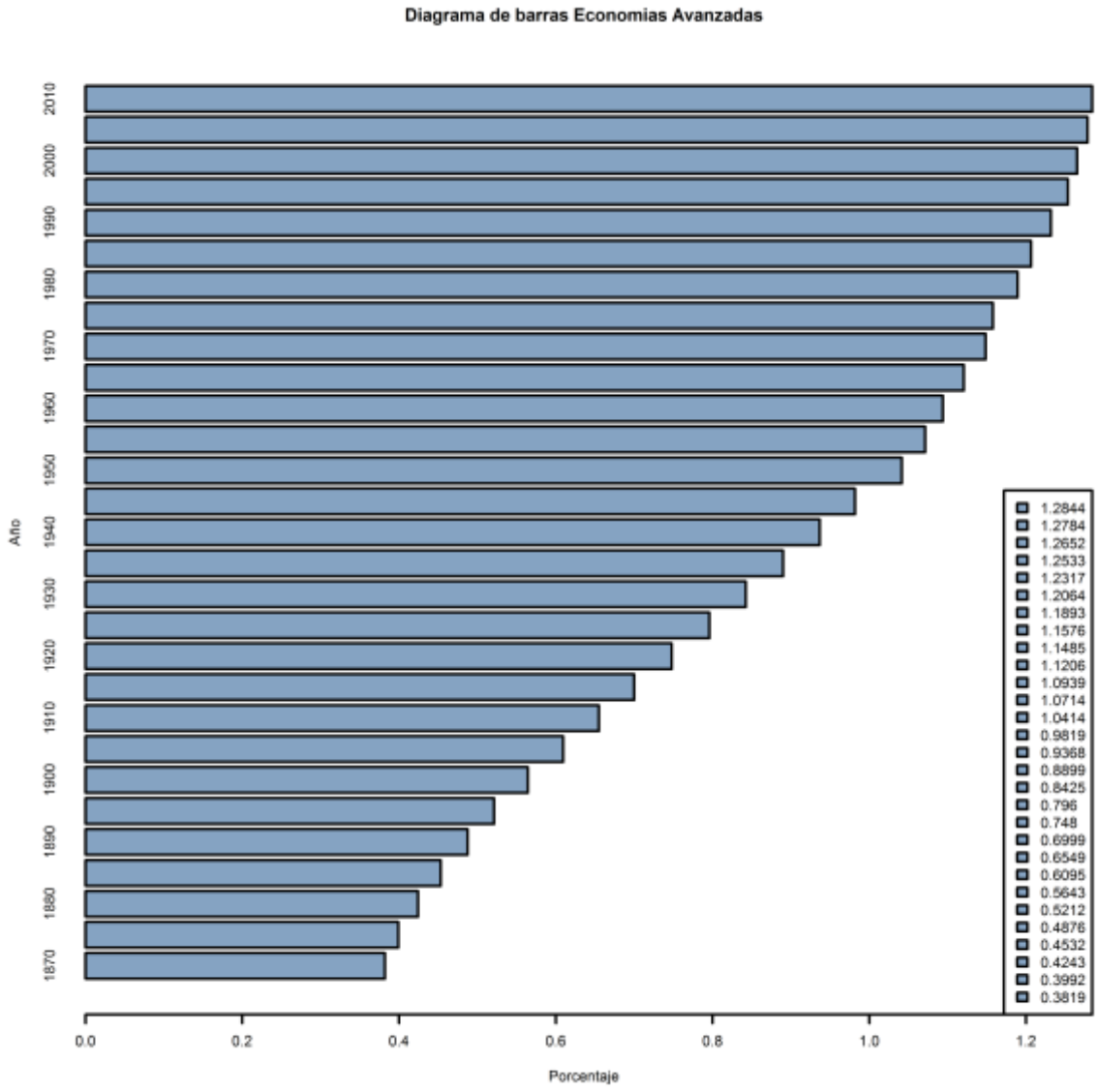


Ilustración 13:Diagramas de barras de Economías Avanzadas

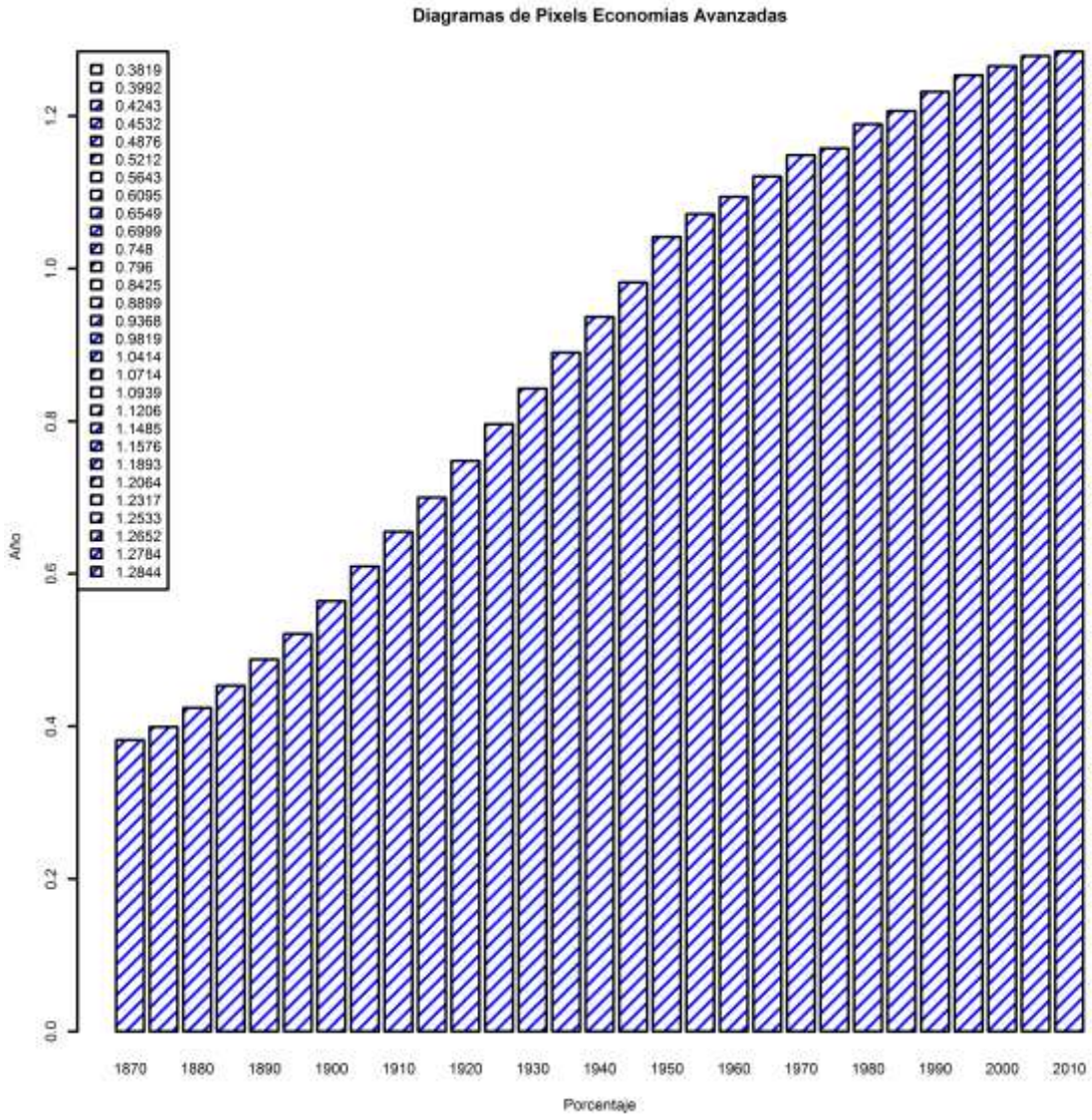


Ilustración 14:Diagramas de Pixeles de Economías Avanzadas

En las representaciones anteriores podemos ver dos técnicas de visualización que aunque categorizadas diferentes son muy similares, observamos en las imágenes que el valor del porcentaje de las personas que estudian aumentan, aunque se ve muy fácil en los ejemplos cuando el número de datos aumenta esta técnicas no se corresponde con los datos así que es mejor elegir otra.

Observamos resultados coherentes para un número bajo de datos, pero cuando los datos aumentan la técnica se convierte en ineficaz por lo que es mejor probar otra.

Ahora probaremos con los diagramas de burbujas, esta técnica de visualización

disponemos de un grafica donde podemos ver el aumento del porcentaje durante los años y si nos colocamos en una burbuja podemos ver el valor exacto en ese año.

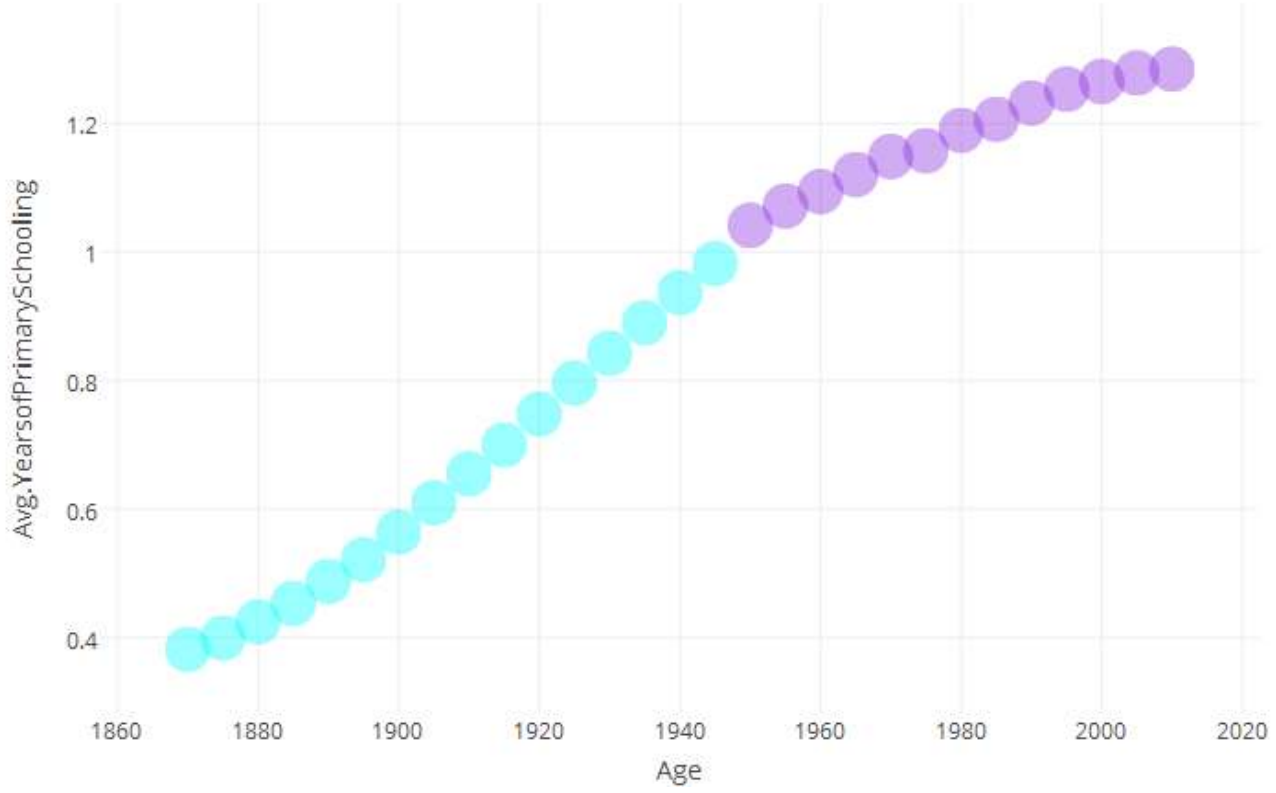


Ilustración 15: Diagrama de burbujas

El grafico anterior es un gráfico que se puede interactuar con él ver los valores año a año seleccionando la burbuja indicada, además de las funciones de un gráfico estándar donde vemos la ascensión de los valores, la ventaja de este tipo de gráficos en lo referente a Big Data es que los valores pueden crecer exponencial sin perder la integridad de los mismos,

El resultado es coherente y podemos decir que se adaptaría a los esperados a una representación Big Data.

A continuación, probaremos las infografías, la técnica a priori mejor se adapta a las características Big Data, puesto que puede incluir diversas técnicas que han obtenido un buen resultado en este caso de uso.

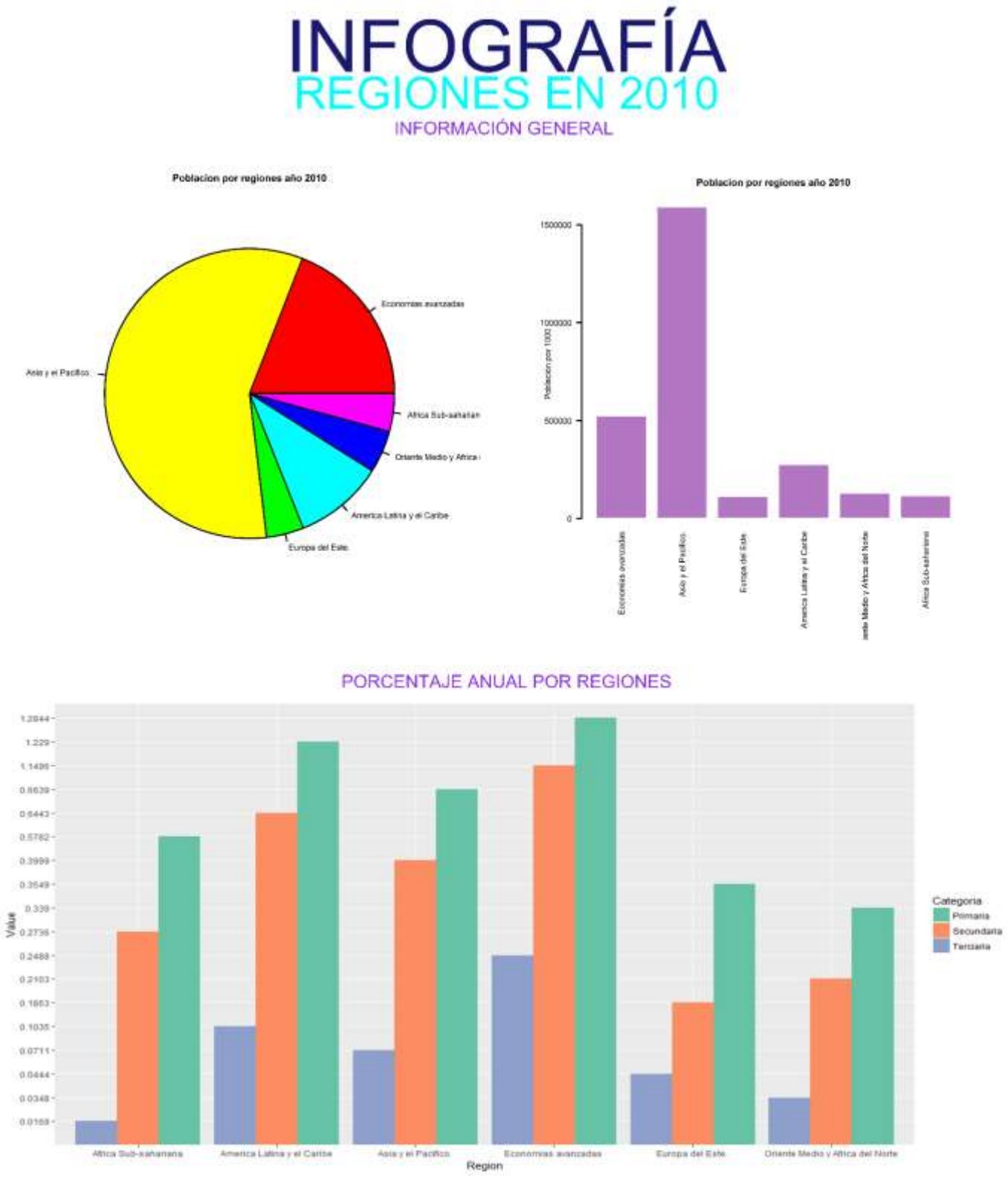


Ilustración 16: Infografía

Como se puede observar en la ilustración anterior hemos unido varias técnicas de forma que se pueden ver rápidamente.

Esta técnica cumple con los requisitos para ser considerada favorable para una correcta visualización de datos en Big Data, representando coherentemente los datos. Pero conlleva la dificultad de saber elegir bien que técnicas emplear.

Seguiremos con los mapas cartográficos, donde podemos representar los valores se toman en cualquier parte del mundo.

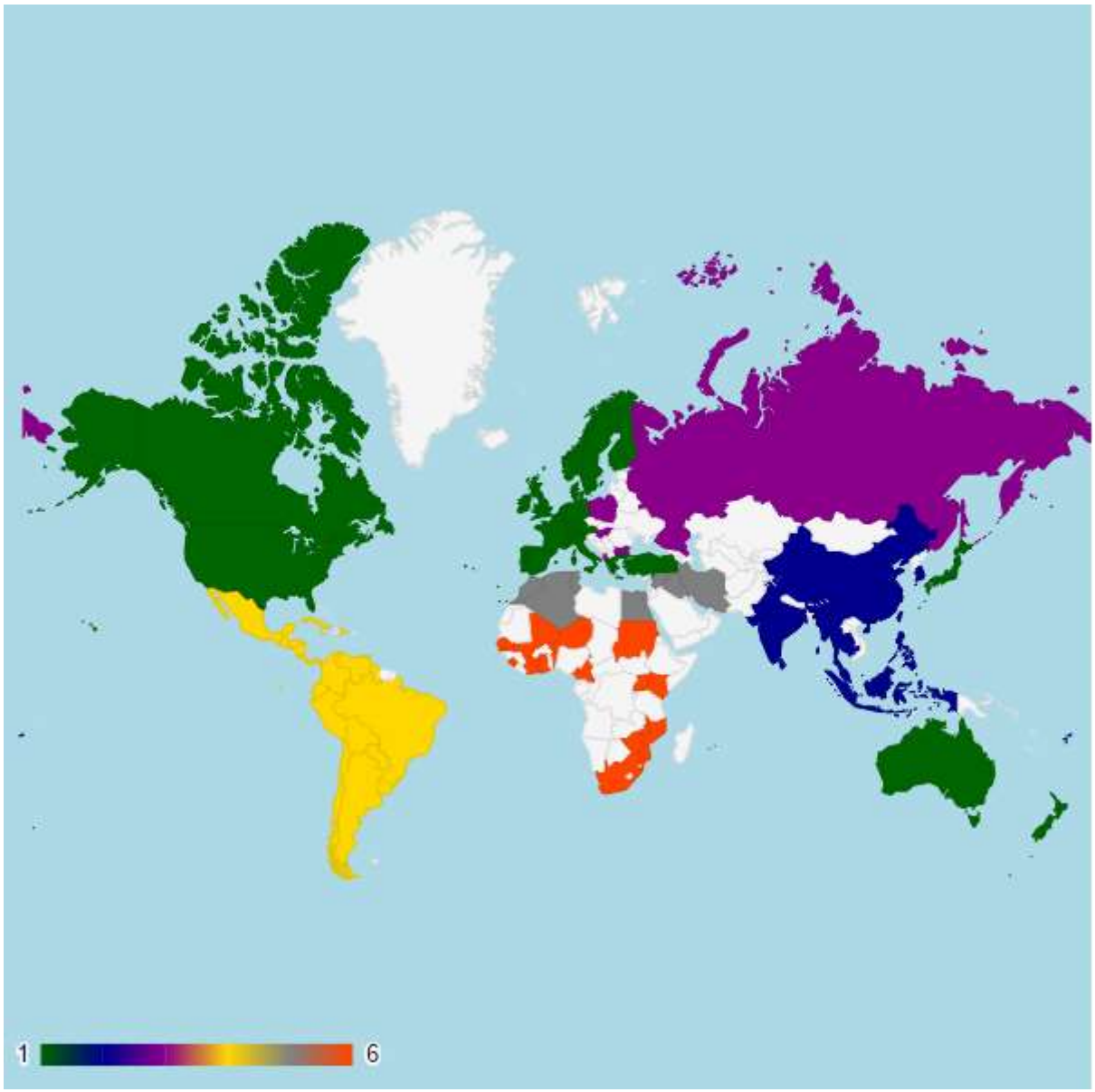


Ilustración 17: Mapa topográfico de las Regiones

En la ilustración anterior podemos ver identificadas las distintas regiones estudiadas diferenciadas por colores, en la esquina inferior derecha podemos ver el código de colores por regiones.

Esta técnica de visualización representa datos coherentes, podemos considerarla buena para Big Data, eso sí es el más trabajoso puesto necesitamos identificar las zonas en el mapa para luego mostrar la información.

Para seguir ahora crearemos una grafico de redes o network, en el cual representamos los datos en forma de grafo no dirigido o de red.

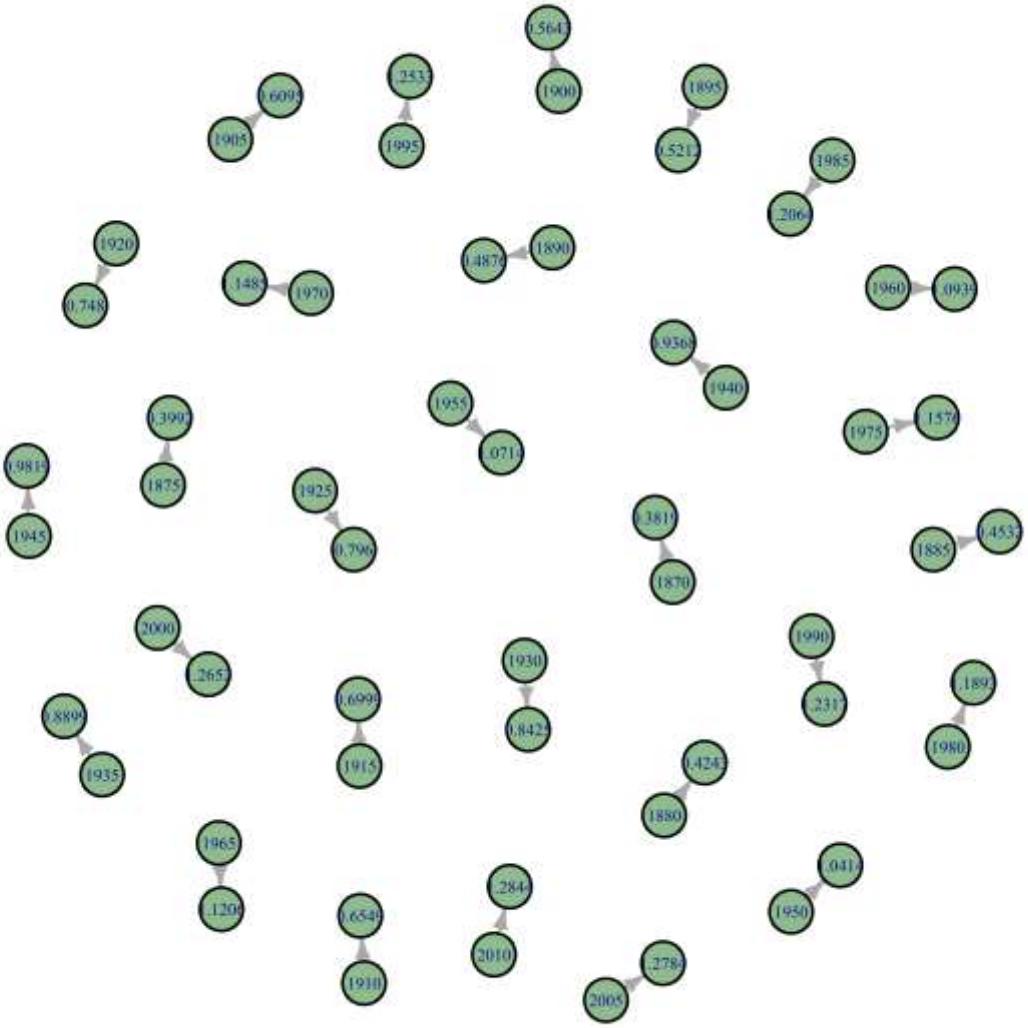


Ilustración 18: Diagrama de Red

En este caso, se ve fácilmente que no es una buena técnica para representar datos en Big Data, puesto que, al tratarse de datos no relacionales, no siempre existe un nexo entre los datos.

Ahora probaremos a realizar un treemap [30] donde representamos los porcentajes anuales de la variable promedio anual escolarizado en primaria, en las cajas que forman el diagrama de árbol.

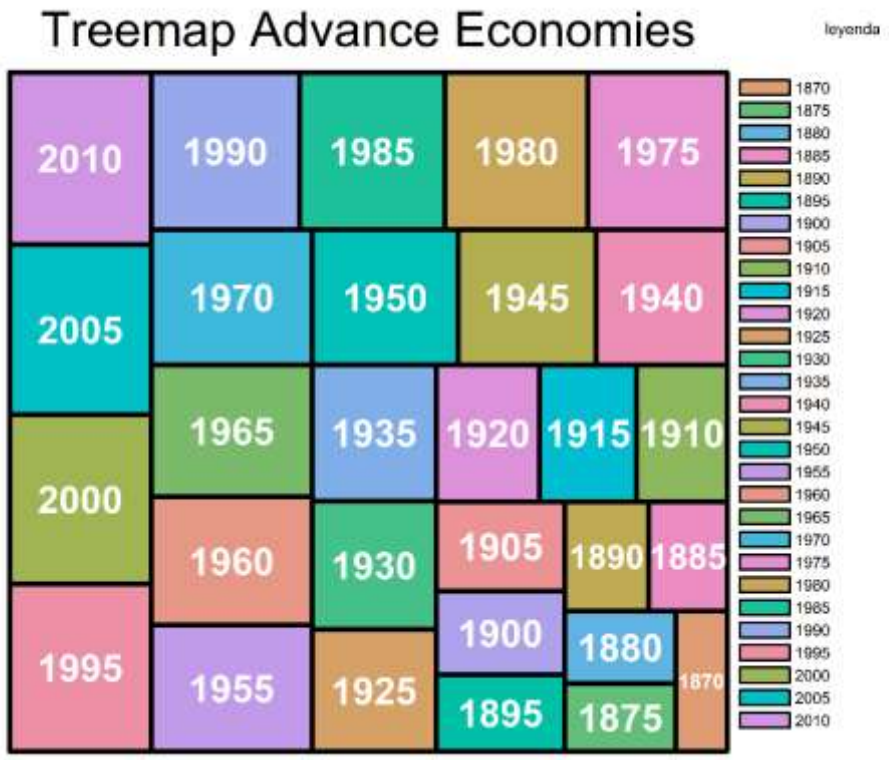


Ilustración 19: Diagrama de árbol

En anterior treemap podemos ver como se representan los años del estudio como rectángulos anidados, el área del rectángulo lo define el valor el porcentaje de personas que estudian en primaria ese año.

Esta técnica muestra contenido coherente y podemos decir que se adapta a los requisitos de Big Data, esta visualización si cumple con las características Bi Data.

Aplicaremos otra técnica de visualización, en este caso probaremos a crear una nube de palabras, por lo que el data set de la población utilizado hasta ahora no nos cumple los requisitos. Por lo tanto, utilizaremos otro conjunto de datos, también de la UNESCO y relacionado con la educación. [8]

Para estudiar el documento utilizaremos el procesamiento de lenguaje natural, con el cual procesamos el documento para obtener la nube de palabras que se repitan más de 50 veces en el documento.

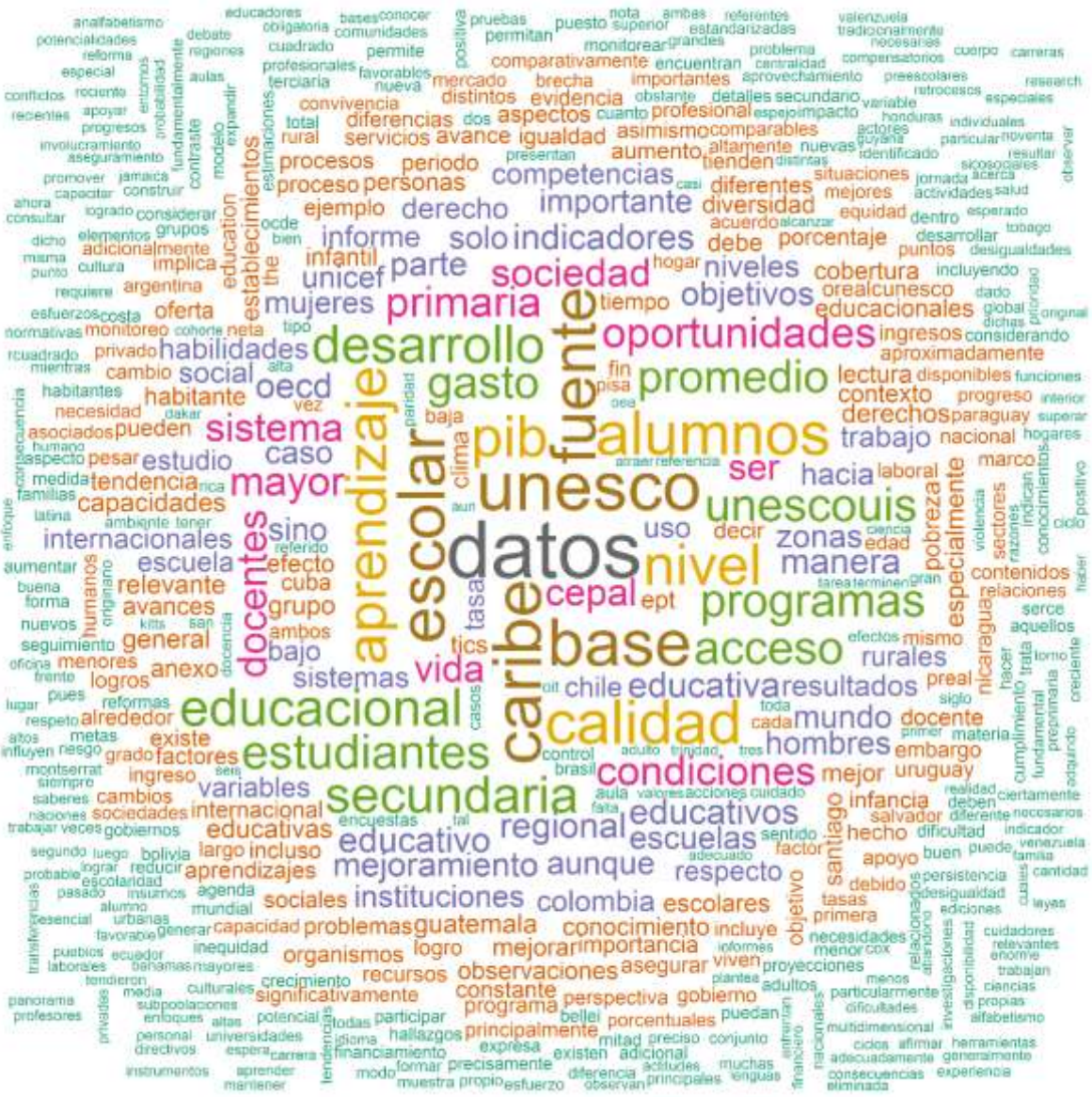


Ilustración 20: Nube de palabras Unesco

En el grafico anterior podemos ver las palabras que más se repiten dentro del documento de educación relacionado con la Unesco, por lo que podemos concluir que las palabras más utilizadas son aprendizaje, conocimiento, educación, entre muchas otras que como se observan aparecen menos.

Concluimos que representa los datos coherentemente y se debe tener en cuenta a la hora de realizar una representación Big Data.

Capítulo 6. Conclusiones y líneas futuras

Como resultado de la investigación Big Data hemos podido descubrir que dentro de la diversidad de técnicas de visualización que existen, no todas se adaptan a las características de Big Data, por lo se ha creado un ranking con las técnicas mejor valoradas según diversos criterios. Tras realizar el caso de estudios aplicando las técnicas mejor clasificadas hemos podido concluir que técnicas se adaptan a Big Data y cuáles no.

También hemos comprobado que la variedad de metodologías existentes no se adapta a nuestros criterios, por lo que hemos concluido realizar nuestra propia metodología para cumplir con nuestros requisitos.

Como líneas futuras tenemos dos posibles ámbitos de estudio, siendo el primero la búsqueda de distintos criterios para clasificar las técnicas de visualización, por lo que se obtiene un nuevo ranking de las técnicas que se adaptarían a Big Data, para luego realizar un caso de estudio y comparar los resultados de las distintas técnicas. El otro ámbito de estudio radicaría en la realización de distintos casos de uso o simplemente representación de la información, puesto que ya se identificaron las técnicas de visualización que mejor se adaptarían a las características Big Data.

Capítulo 7. Summary and Conclusions

As a result of the Big Data research, we have been able to discover that, within the diversity of visualization techniques that exist, not all are adapted to the characteristics of Big Data. Therefore, a ranking has been created with the best valued techniques according to various criteria. After carrying out the case studies, applying the best classified techniques, we have been able to conclude which techniques are adapted to Big Data and which are not.

We have also verified that the variety of existing methodologies does not fit our criteria, so we have concluded our own methodology to meet our needs.

As future lines, we have two possible fields of study, being the first the search of different criteria to classify the visualization techniques, so that a new ranking of the techniques that would adapt to Big Data is obtained, to then make a case study and compare the results of the different techniques. The other field of study would be the realization of different use cases or, simply, representating the information, since the visualization techniques that would better adapt to the Big Data characteristics were already identified.

Capítulo 8. Presupuesto

El presupuesto de este proyecto está basado en las horas de trabajo que se han empleado para realizar el trabajo, puesto que nuestro proyecto es de investigación, no ha sido necesario adquirir ningún tipo de instrumental, hardware privativo ni licencias.

Horas de trabajo	Descripción	Costo
160 horas	Estudiar el estado del arte relacionado con Big Data y visualización.	**160 * 9.37 = 1499,2**
120 horas	Buscar, definir y localizar las diferentes técnicas de visualización y técnicas predictivas.	**120 * 9.37 = 1124,4**
125 horas	Buscar y comparar metodologías existentes, para visualizar los datos con técnicas predictivas de Big Data.	**125 * 9.37 = 1171,25**
45 horas	Proponer metodología acorde a las necesidades encontradas.	**45 * 9.37 = 421,65**
160 horas	Realizar caso de uso.	**160 * 9.37 = 1499,2**
82 horas	Crear documentación técnica y la memoria del proyecto.	**82 * 9.37 = 768,34**
	Total:	**6783,88**

Tabla 4: Presupuesto

El cálculo del precio por hora es obtenido suponiendo un sueldo mensual de 1500 euros, dividido entre 40 horas semanales durante 4 semanas obteniendo un precio por hora de 9.37€

Referencias

[1] B. Marr, Big Data: Using SMART big data, analytics and metrics to make better decisions and improve performance, John Wiley & Sons, 2015.

[2] C. Bielza, A. Salmeron, A. Alonso-Betanzos, J. I. Hidalgo, L. Martínez, A. T. Lora, E. Corchado y J. M. Corchado, Advances in Artificial Intelligence: 15th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2013, Madrid, September 17-20, 2013, Proceedings, vol. 8109, Springer, 2013.

[3] Mayer-Schönberger, V., & Cukier, K, Big data: la revolución de los datos masivos., Turner, 2013.

[4] A. Kirk, Data Visualization: a successful design process, Packt Publishing Ltd, 2012.

[5] R. A. Española, *Visualizar*, 2014.

[6] R. A. Española, *Educacion*, 2014.

[7] Unesco, *Datos educacion*, 2010.

[8] Unesco, *Datos educacion Texto*, 2015.

[9] G. A. &. I. Elcoro, *Introducción a Big Data*, 2015.

[10] D. C. N. &. B. M. Batista, *Big Data Análisis de Datos*, 2013.

[11] F. B. Data, *Fundación Big Data, glosario Big Data*, 2017.

[12] F. Castanedo, *Una visión de las técnicas y métodos de análisis en entornos Big Data utilizando tecnologías de última generación*, 2015.

[13] L. M. Padua, «Comparación interactiva de modelos de minería de datos utilizando técnicas de visualización,» 2014.

[14] OA.Usun, *Visualización de la información: la base del conocimiento enlazando ideas de forma gráfica,*, 2012.

[15] M. T. S. Martín, «Apuntes sobre la información visual en la formación online de las futuras generaciones,» *RELADA-Revista Electrónica de ADA-Madrid*, vol. 2, 2009.

[16] mpc, *5 herramientas esenciales para hacer infografías*, 2012.

[17] RAE, «Real Academia Española Metodología,» [En línea]. Available: http://dle.rae.es/srv/search?m=30&w=metodología.

[18] M. R. a. R. D. G. a. T. S. L. a. B. F. Martinez-Torres, *Metodologías de Análisis de los Big Data en las Plataformas Educativas*, 2015.

[19] k. d. p. s. suarez, «metodobayes.blogspot.com.es,» [En línea]. Available: http://metodobayes.blogspot.com.es/2013/05/metodo-bayes.html.

[20] U. n. i. d. organization, Unido technology foresight manual Volume 1 Organization and Methods, Viena: United nations industrial development organization, 2005.

[21] Gallardo, «Universidad de Granada,» [En línea]. Available: http://www.ugr.es/~gallardo/pdf/cluster-3.pdf.

[22] F. J. Romero-Campero, «Universidad de Sevilla,» [En línea]. Available: https://www.cs.us.es/~fran/curso_unia/clustering.html.

[23] P. C., «http://www.dataprix.com,» [En línea]. Available: http://www.dataprix.com/blog-it/business-intelligence/mineria-datos/data-mining-basico-correlaciones-regresiones-mercado-valores-excel.

[24] O. R. Rojas, «http://www.oldemarrodriguez.com/,» [En línea]. Available: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Presentación_-_Conceptos_Básicos.41132532.pdf.

[25] CICLing, «http://www.cicling.org,» [En línea]. Available: http://www.cicling.org/ampln/NLP.htm.

[26] vicomtech, «http://www.vicomtech.org,» [En línea]. Available: http://www.vicomtech.org/t4/e11/procesamiento-del-lenguaje-natural.

[27] W. H. Inmon, Building the data warehouse, John wiley & sons, 2005.

[28] S. Paniagua, «sorayapaniagua,» [En línea]. Available: http://www.sorayapaniagua.com/2011/11/01/la-ciencia-de-los-datos-bdii/.

[29] N. Zumel, J. Mount y J. Porzak, Practical data science with R, Manning, 2014.

[30] T. S. Inc., «Tibco,» [En línea]. Available: https://docs.tibco.com/pub/spotfire_web_player/6.0.0-november-2013/es-ES/WebHelp/GUID-F3F4ABDF-8418-42D3-A1C4-60B7A8121C75.html.

[31] A.-A. Asaad, «Blog de R,» [En línea]. Available: https://www.r-bloggers.com/r-how-to-layout-and-design-an-infographic/.

[32] Comunidad, Yan Holtz, «R graph gallery,» [En línea]. Available: http://www.r-graph-gallery.com.

[33] E. F. Glynn, «Earl F Glynn Github,» [En línea]. Available: http://earlglynn.github.io/RNotes/package/gplots/heatmap2.html.

[34] C. R, «R documentation,» 2017. [En línea]. Available: rdocumentation.org. [Último acceso: 2017].

[35] R. p. C., *cran.r-project.org*, 2014.

[36] J. Serrano-Cobos, «Big data y analítica web. Estudiar las corrientes y pescar en un océano de datos,» *El profesional de la información*, vol. 23, pp. 561-565, 2014.

[37] O. –. D. M. F. &. Fun, *Orange – aprendizaje automático,* 2017.

[38] R. S. J. D. Baker y K. Yacef, «The state of educational data mining in 2009: A review and future visions,» *JEDM-Journal of Educational Data Mining*, vol. 1, pp. 3-17, 2009.

[39] A. Arcavi, «The role of visual representations in the learning of mathematics,» *Educational studies in mathematics*, vol. 52, pp. 215-241, 2003.

[40] R. Nugent, N. Dean y E. Ayers, «Skill set profile clustering: the empty K-means algorithm with automatic specification of starting cluster centers,» de *Educational Data Mining 2010,* 2010.

[41] A. Hadoop, *Apache Hadoop,* 2017.

[42] M. P. Estes, *Apache spark: qué es y cómo funciona,* 2015.

Apéndice A. Gráficas en R

A.1. Algoritmo Incorporación dataset para crear diagramas.

```
##############################################################################################
#
# dataProcessing.r
#
##############################################################################################
#
# AUTOR: YERAY PEREZ PERAZA
#
#
# FECHA: 19/04/2017
#
#
# DESCRIPCIÓN: Fichero realizado en R, para la incorporación de datos para su
   posterior procesamiento hasta lograr, crear un diagrama o grafica representado la
   información almacenada en los datos.
#
#
AdvancesEconomies <-
  read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/ae.csv",
             header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
AsiaAndThePacific <-
  read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/atp.csv",
             header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
EasternEurope <-
  read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/ee.csv",
             header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
LatinAmericaAndTheCaribean <-
  read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/latc.csv",
             header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
MiddleEastAndNorthAfrica <-
  read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/mena.csv",
```

```
            header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
SubSaharanAfrica<-
  read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/mena.csv",
            header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)

##############################################################################################
```

A.2. Algoritmos para la creación de graficas

```
##############################################################################################
#
 # dataSetPoblacion.r
#
##############################################################################################
#
# AUTOR: YERAY PEREZ PERAZA
#
#
# FECHA: 31/05/2017
#
#


### Mapas de calor
#### Ejemplo con Advance Economies
library(gplots)
library(RColorBrewer)
AdvancesEconomies <-
  read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/ae.csv",
            header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
# creamos estructuras de los datos.
rnames <- AdvancesEconomies[,1]
data <- data.matrix(AdvancesEconomies[,4:ncol(AdvancesEconomies)])
rownames(data) <- rnames
#definimos la paleta de colores
```

```
my_palette <- colorRampPalette(c("#0A0A2A", "#2E9AFE", "#A9F5F2"))(n = 299)
# identificamos los valores límites de cada color.
col_breaks = c(seq(-1,0,length=100),  # for primary blue
               seq(0.01,0.8,length=100),           # for secundary blue
               seq(0.81,1,length=100))             # for terciary blue

#creamos el archivo png para guardar el grafico
png(
  "C:/r/TFG/HeatMaps_in_r_ae1.png",
  width     = 3.25,
  height    = 3.25,
  units     = "in",
  res       = 1200,
  pointsize = 4
)

Creamos el mapa de calor
heatmap.2(data,
          main = "HeatMaps Of Advances Economies", # nombre del grafico
          density.info="none",  #impedimos que en la leyenda se vea la densidad
          trace="none",         # eliminamos del grafico la progresión de los valores
          margins =c(11,5),     # margen inferior y superior
          col=my_palette,       # asignamos el color que creamos anteriormente
          breaks=col_breaks,    # indicamos los límites del color
          keysize = 1,          # tamaño de la leyenda de color
          labCol= c("Primary","Secondary","Tertiary"), # nombre de las columas
          scale="column",  #ordenamos por el nombre de las columnas
          dendrogram="both", # eliminamos el dendograma
          lwid=c(0.55,1.33)  # limitamos el acho de las columnas
)
dev.off()

  # Todos los mapas de calor se crean igual simplemente cambiamos el origen de los
datos y el nombre del fichero png donde guardamos los resultados.


  # Diagrama de puntos ------------------------------------------------------
```

```
#### Ejemplo con Advance Economies
library(gplots)
library(RColorBrewer)
AdvancesEconomies <-
  read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/ae.csv",
             header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
png(
  "C:/r/TFG/DotDiagram_in_r_ae1.png",
  width     = 7.25,
  height    = 7.25,
  units     = "in",
  res       = 1200,
  pointsize = 4
)

xyplot(Avg.YearsofPrimarySchooling + Avg.YearsofSecondarySchooling +
         Avg.YearsofTertiary.Schooling ~ Age, type=c("p", "l"), pch=16,
       auto.key=list(border=TRUE), par.settings=simpleTheme(pch=16),
       scales=list(x=list(relation='free'), y=list(relation='free')),
data=AdvancesEconomies,  xlab="Age", ylab="Average", main="Dot diagram Advances
Economies")
# en la instrucción anterior seleccionamos las variables, indicamos que el grafico es
de líneas y puntos, que las escalas son libres, es decir, se crean según los datos,
indicamos el origen de los datos y finalmente los nombres de los ejes y del prpio
grafico
dev.off()

### Diagramas de volumen
#### Diagrama de barras
library(gplots)
ae <- read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/ae.csv",
                 header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
# creamos la imagen png
png(
  "C:/r/TFG/BARPLOT_AE.png",
  width     = 4.25,
  height    = 4.25,
```

```
  units     = "in",
  res       = 1200,
  pointsize = 4
)
#creamos estructuras de datos
vector1 = c(ae[,4])
names(vector1)=c(ae[,1])
barplot(vector1,col=rgb(0.2,0.4,0.6,0.6) , border="black" , horiz=T ,
 main = "Diagrama de barras Economias Avanzadas" , xlab = "Porcentaje", ylab = "Año",
legend = colnames(vector1), legend.text = c(ae[,4]),  args.legend = list(x =
"bottomright"))
 # Con la instrucción creamos el diagrama de barras, indicamos los datos, el color de
las barras, la orientación de las barras, las etiquetas, y los datos para a leyenda.
dev.off()



##### Pixel Bars
library(gplots)
ae <- read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/ae.csv",
                 header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
# creamos la imagen png
png(
  "C:/r/TFG/Pixel_Bars_AE.png",
  width     = 4.25,
  height    = 4.25,
  units     = "in",
  res       = 1200,
  pointsize = 4
)
#creamos estructuras de datos
vector1 = c(ae[,4])
names(vector1)=c(ae[,1])
barplot(vector1,col="blue" , border="black"  , main = "Diagramas de Pixels Economias
Avanzadas" ,  xlab = "Porcentaje", ylab = "Año" , legend = colnames(vector1),
legend.text = c(ae[,4]),    args.legend = list(x = "topleft"), density=25)
# Con la instrucción creamos el diagrama de pixeles primero indicamos los datos, el
color de las barras, la orientación de las barras, las etiquetas, los datos para a
leyenda y indicamos que sea de pixeles con una densidad de especifica.
```

```
dev.off()

# nube de palabras ----------------------------------------------------------

#indicamos los paquetes que vamos a usar
library(tm)
library(wordcloud)
library(lsa)



#cargamos los ficheros a usar
txt0 <- readLines("C:/Users/Yery/Desktop/UNI-15-16/16-
    17/TFG/dataSet/educacion_unesco.txt",encoding = "UTF-8")
txt0 = iconv(txt0,to="ASCII//TRANSALIT")
#creamos nuevo corpus
corpus_new <- Corpus(VectorSource(txt0))
#eliminar espacios, minúsculas y puntuación
d_new <- tm_map(corpus_new,content_transformer(tolower))
d_new <- tm_map(d_new,stripWhitespace)
d_new <- tm_map(d_new,removePunctuation)
d_new <-tm_map(d_new, removeNumbers)
#palabras vacías
d_new <- tm_map(d_new,removeWords,stopwords("spanish"))
#nueva matriz de elementos
tdm_new <- TermDocumentMatrix(d_new)
m_new = as.matrix(tdm_new)
wf <- sort(rowSums(m_new),decreasing=TRUE)
dm <- data.frame(word = names(wf), freq=wf)
#indicamos que la renuencia es 50 apariciones
findFreqTerms(tdm_new, lowfreq=50)
#creamos la imagen png
png(
  "C:/r/TFG/nube_unesco.png",
  width     = 3.25,
  height    = 3.25,
```

```
  units     = "in",
  res       = 1200,
  pointsize = 4
)
#creamos la nube de palabras con los valores anteriores
wordcloud(dm$word, dm$freq, random.order=FALSE, colors=brewer.pal(8,"Dark2"))

dev.off()

# treeMap ----------------------------------------------------------------

library(treemap)
ae <- read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/ae.csv",
                 header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
# creamos la estructura de los datos
data1=data.frame(group=c(ae[,4]) , value=c(ae[,1]))
png(
  "C:/r/TFG/treemap_ae.png",
  width     = 4.25,
  height    = 3.25,
  units     = "in",
  res       = 1200,
  pointsize = 4
)
treemap(data1,
        index="value",
        vSize="group",
        type="index",
        fontcolor.labels=c("white","orange"),
        title="Treemap Advance Economies" ,
        title.legend = "leyenda",
        position.legend = "right",
        fontsize.legend = 4.5

)
# Con la instrucción anterior indicamos el origen de los datos, los colores a usar,
```

el título y las características de la leyenda.

```
dev.off()


# grafico de burbujas ------------------------------------------------


library(plotly)
ae <- read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/ae.csv",
                 header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
head(ae)
plot_ly(ae, x = ~Age, y = ~Avg.YearsofPrimarySchooling, type="scatter", mode =
"markers" ,     marker=list(
color=ifelse(ae$Avg.YearsofPrimarySchooling>1,"aquamarine","coral") , opacity=0.4 ,
size=25) )
```

En las instrucciones anteriores creamos el diagramas de burbujas indicando las variables a representar, que el tipo de grafico es de burbujas y el filtro de color según el valor de la variable y por último el tamaño de la burbuja.

```
# network ------------------------------------------------------------------


library(igraph)
ae <- read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/ae.csv",
                 header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
png(
  "C:/r/TFG/network.png",
  width     = 4.25,
  height    = 4.25,
  units     = "in",
  res       = 1200,
  pointsize = 4
)
# creamos la estructura de los datos
nodos=data.frame(
  source=c(ae[,1]),
  target=c(ae[,4])
)
network=graph_from_data_frame(d=nodos, directed=T)
deg=degree(network, mode="total")
```

```
   plot(network, vertex.size=deg*9, vertex.color="DarkSeaGreen")

   # creamos el diagrama de red, para ello primero indicamos la matriz, indicamos los
grados de la red y por ultimo creamos la red, para ello indicamos la matriz de nodos,
el tamaño de los vértices y el color de los mismos.

   dev.off()




   # mapa ----------------------------------------------------------------

library(googleVis)
#cargamos los datos de las regiones
all <- read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/datos4.csv",
header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
#paises pertenecientes a la UNESCO
  Country =
c("Australia","Austria","Belgium","Canada","Denmark","Finland","France","Germany","Gree
ce","Ireland","Italy","Japan","Luxembourg","Netherlands","New
Zealand","Norway","Portugal" , "Spain","Sweden","Switzerland","Turkey","USA","United
Kingdom","Cambodia","China","Fiji","Hong Kong, China","India","Indonesia","Malaysia",
"Myanmar","Philippines","Republic of Korea", "Sri Lanka", "Taiwan", "Thailand",
"Albania","Bulgaria", "Hungary","Poland"," Russian Federation",
"Argentina","Barbados","Bolivia", "Brazil", "Chile", "Colombia","Costa Rica"," Cuba","
Dominican Rep.","Ecuador","El Salvador","Guatemala","Guyana","Honduras",
"Jamaica","Mexico", "Nicaragua", "Panama","Paraguay","Peru", " Trinidad and
Tobago","Uruguay","Venezuela", "Algeria","Egypt","Iran", "Iraq", "Malta", "Morocco",
"Syria", "Tunisia","Benin","Cameroon", "Cote dIvoire", "Ghana","Kenya", "Malawi",
"Mali", "Mauritius","Mozambique", "Niger","Reunion" , "Senegal","Sierra Leone"," South
Africa",  "Sudan", "Uganda", "Zimbabwe")

#identificamos los paises de cada region

  Profit = c( seq(1, 1, length=23), seq(2,2,length=13), seq(3,3,length=5),
seq(4,4,length=23), seq(5,5,length=8),seq(6,6,length=17))

#estructura de datos que une los paises con el identificador de la región

paises=data.frame(Country, Profit)

#Indicamos las regiones
Region=c("Economias avanzadas", "Asia y el Pacifico.", "Europa del Este.","America
Latina y el Caribe", "Oriente Medio y Africa del Norte", "Africa Sub-sahariana")

#Indicamos el color de cada region

Color= c("DarkGreen", "DarkBlue","DarkMagenta","Gold","Grey","OrangeRed")

#indicamos el año de los datos

Año=c("2010")
```

```
#mostramos los datos de cada categoría

Primaria=c(paste0(all[29,5]),paste0(all[58,5]),paste0(all[87,5]),paste0(all[116,5]),pas
te0(all[145,5]),paste0(all[174,5]))
Secundaria=c(paste0(all[29,6]),paste0(all[58,6]),paste0(all[87,6]),paste0(all[116,6]),p
aste0(all[145,6]),paste0(all[174,6]))
Terciaria=c(paste0(all[29,7]),paste0(all[58,7]),paste0(all[87,7]),paste0(all[116,7]),pa
ste0(all[145,7]),paste0(all[174,7]))
#creamos la estructura de los datos con las regiones, el año y los datos de cada
categoría

leyenda <- data.frame(Region,Color,Año, Primaria,Secundaria,Terciaria)


#cremos el mapa con los datos

Geo=gvisGeoChart(paises, locationvar="Country",
                 colorvar="Profit",
                 options=list(colors="['#006400', '#00008B', '#8B008B', '#FFD700',
'#808080', '#FF4500']", width=520, height=520,backgroundColor="lightblue"))


#mostramos el resultado

plot(Geo)


   #Infografía ----------------------------------------------------------------


   library(ggplot2)
   library(lattice)
   library(graphics)
   library(car)
   library(grid)
   library(plotrix)


   all <- read.table("C:/Users/Yery/Desktop/UNI-15-16/16-17/TFG/dataSet/datos4.csv",
                     header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
   #creamos los gráficos para la infografía
   maleta <- readPNG(system.file("img", "mochilaCole.png", package="png"), TRUE)


   slices <- c(all[29,3],all[58,3],all[87,3],all[116,3],all[145,3],all[174,3])
   Region=c("Economias avanzadas", "Asia y el Pacifico.", "Europa del Este.","America
Latina y el Caribe",  "Oriente Medio y Africa del Norte", "Africa Sub-sahariana")
   g9 <-pie(slices, labels = Region, main="Poblacion por regiones año 2010",
```

```
col=rainbow(length(Region)))


   vector.all =c(all[29,3],all[58,3],all[87,3],all[116,3],all[145,3],all[174,3])

   names(vector.all)=c("Economias avanzadas", "Asia y el Pacifico.", "Europa del
Este.","America Latina y el Caribe","Oriente Medio y Africa del Norte", "Africa Sub-
sahariana")


   par(mar=c(12,5,5,5))

   diagrama <-barplot(vector.all, col=rgb(0.5,0.1,0.6,0.6), ylab="Población por 1000",
main="Poblacion por regiones año 2010", las=2, border=F)

   Region=c("Economias avanzadas", "Asia y el Pacifico.", "Europa del Este.","America
Latina y el Caribe", "Oriente Medio y Africa del Norte", "Africa Sub-sahariana")

   Primaria=c(paste0(all[29,5]),paste0(all[58,5]),paste0(all[87,5]),paste0(all[116,5]),p
aste0(all[145,5]),paste0(all[174,5]))

   Secundaria=c(paste0(all[29,6]),paste0(all[58,6]),paste0(all[87,6]),paste0(all[116,6])
,paste0(all[145,6]),paste0(all[174,6]))

   Terciaria=c(paste0(all[29,7]),paste0(all[58,7]),paste0(all[87,7]),paste0(all[116,7]),
paste0(all[145,7]),paste0(all[174,7]))

   Categoria=c("Primaria","Primaria","Primaria","Primaria","Primaria","Primaria","Secund
aria","Secundaria","Secundaria","Secundaria","Secundaria","Secundaria","Terciaria","Ter
ciaria","Terciaria","Terciaria","Terciaria","Terciaria")

   Value=vector()

   Value=c(Value,Primaria)

   Value=c(Value,Secundaria)

   Value=c(Value,Terciaria)

   data3 <- data.frame(Region, Categoria,Value)


   g10<-ggplot(data3, aes(fill=Categoria, y=Value, x=Region)) +
     geom_bar(position="dodge", stat="identity")+
     scale_fill_brewer(palette = "Set2")


   # crear imagen ------------------------------------------------------------


   png(
     "C:/r/TFG/infografia.png",
     width     = 6.25,
     height    = 7.25,
     units     = "in",
     res       = 1200,
```

```
    pointsize = 4
  )


  grid.newpage()
  pushViewport(viewport(layout = grid.layout(4, 3)))
  grid.text("INFOGRAFÍA", y = unit(0.96, "npc"), x = unit(0.5, "npc"),gp = gpar(col =
"#191970", cex = 8))
  grid.text("REGIONES EN 2010", y = unit(0.92, "npc"), gp = gpar( col = "#00FFFF", cex
= 5))
  grid.text("INFORMACIÓN GENERAL ", y = unit(0.89, "npc"), gp = gpar( col = "#8A2BE2",
cex = 2))
  #print(maleta, vp = vplayout(2, 1:1))
  print(g9, vp = vplayout(2, 1:2))
  print(diagrama, vp = vplayout(2, 2:3))
  grid.text("PORCENTAJE ANUAL POR REGIONES", y = unit(0.42, "npc"), gp = gpar( col =
"#8A2BE2", cex = 2))
  print(g10, vp = vplayout(4, 1:3))
  dev.off()


  # Fuentes de los códigos: [31] [32] [33] [34]
  ##############################################################################################
```