ITESO

Trabajadores asegurados por actividad económica para todas las entidades federativas

Análisis de similitud entre variables dentro de la base de datos del gobierno

Víctor Ramos y Eduardo Castillo

3 Marzo 2017

Introducción:

La finalidad de este trabajo es aplicar los temas vistos en la clase de Ciencia de Datos e Inteligencia de Negocios en una base de datos real con todas las implicaciones que eso conlleva. Para esto la base de datos utilizada fue "Trabajadores asegurados por actividad económica para todas las entidades federativas y sus municipios" obtenida desde la página https://datos.jalisco.gob.mx/.

Limpieza y extracción de la información estadística:

Se eligió la base de datos del 2015 ya que ésta contenía información más reciente. Primero fue necesario verificar que los datos estuvieran bien ordenados ya que algunas descripciones de las actividades laborales tomaban dos columnas y no una, lo que desfasaba los datos en las filas donde ocurría eso.

Una vez que los datos estuvieron bien ordenados se quitaron las columnas que contenían información repetida ya que varias de las columnas eran sumas de otras. Por lo tanto, la información que se podría obtener de éstas era redundante.

Finalmente se decidió agrupar todos los municipios de los estados de acuerdo al sector laboral para que el análisis estadístico fuera entre estados y no municipios, respetando todas las demás variables como mes y sub rubro laboral.

Dado que se quiso añadir como variable a analizar el mes se transformó la variable categórica que contenía dicha información en varias variables "dummy" donde 1 representa el mes en la columna que coincida y 0 para todos las demás columnas de los demás meses.

Los resultados para los reportes de calidad de los datos y utilizando la función "describe" son los siguientes:

Index	Data type	missing values	present values	unique values	minimum values	maximun values
Mes	object	0	469543	6	Abril	Mayo
Entidad_Fede…	object	0	469543	32	AGUASCALIENT…	ZACATECAS
Municipio	object	0	469543	1.91e+03	ABALA	ZUMPANGO
Division_de_…	object	0	469543	8	Agricultura, Ganaderia, S…	Transportes y Comunicacion…
Trabajadores…	int64	0	469543	7.1e+03	0	199851
Trabajadores…	int64	0	469543	2.57e+03	0	22671
Trabajadores…	int64	0	469543	6.5e+03	0	180464
Trabajadores…	int64	0	469543	818	0	13262
Trabajadores…	int64	0	469543	7.17e+03	1	199851

Reporte de calidad de datos

Index	count	unique	top	freq
Mes	469543	6	Junio	78722
Entidad_Fede…	469543	32	JALISCO	44320
Municipio	469543	1913	GUADALAJARA	6087
Division_de_…	469543	8	Comercio	157526

Resultado función describe

Index	Trabajadores_Asegurados
Agricultura, Ganaderia, Silvicultura, Pesca y Caza	3502720
Comercio	21192366
Ind ElÚctrica y Captaci¥n y Suministro de Agua Potable	870979
Industria de la Construcci¥n	8827989
Industrias Extractivas	776809
Industrias de la Transformaci¥n	27856970
Servicios	36519285
Transportes y Comunicaciones	5596983

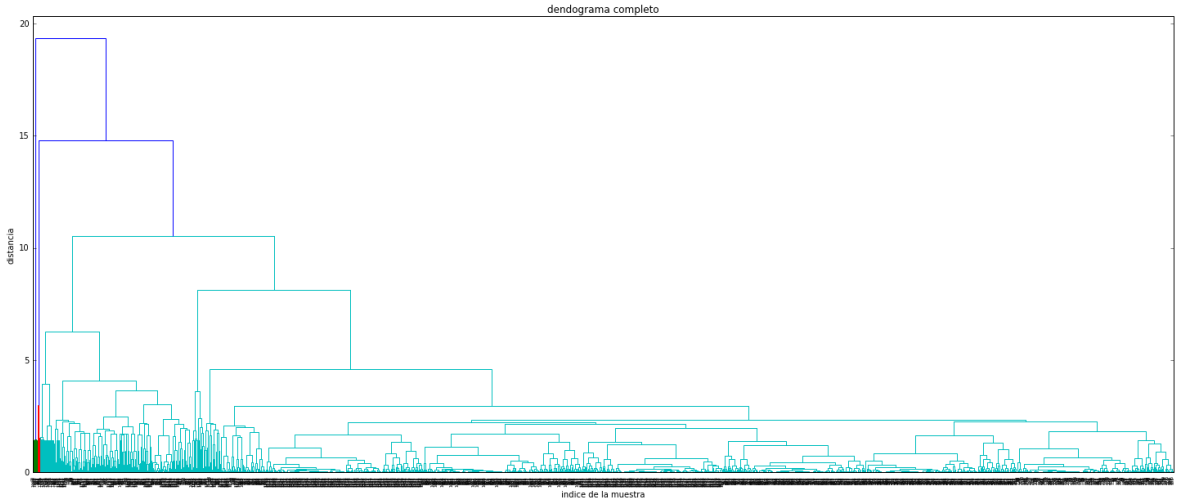
Número de asegurados por sector laboral

Index	Trabajadores_Asegurados
AGUASCALIENT…	1572104
BAJA CALIFORNIA	4472325
BAJA CALIFORNIA S…	831506
CAMPECHE	882187
CHIAPAS	1289882
CHIHUAHUA	4642851
COAHUILA	4127367
COLIMA	720231
DISTRITO FEDERAL	18417496
DURANGO	1327298
ESTADO DE MEXICO	8315831
GUANAJUATO	4934431
GUERRERO	918815
HIDALGO	1220946

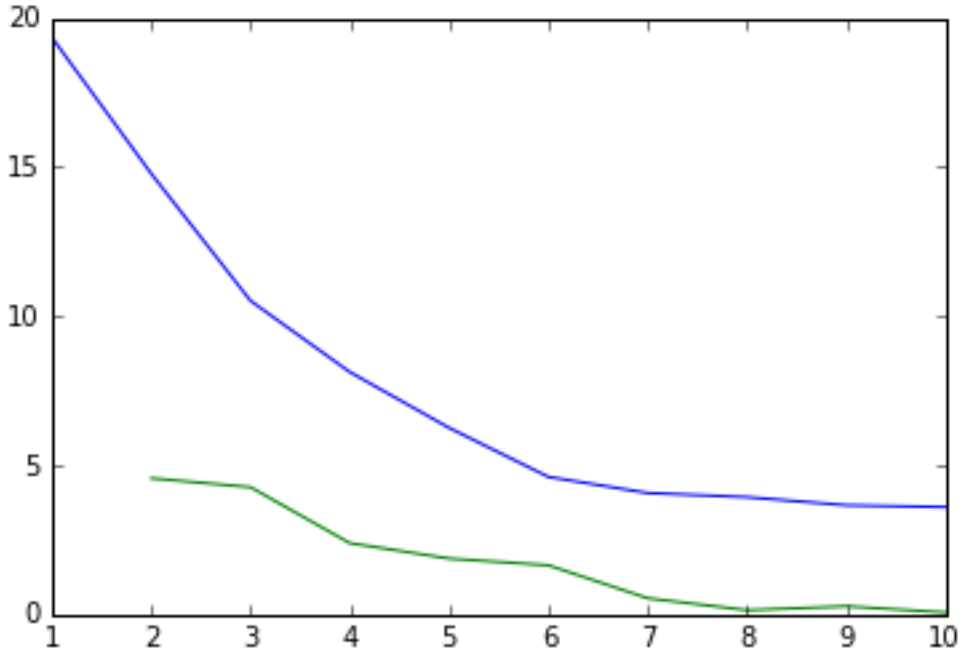
Número de asegurados por estado

Agrupamiento de datos:

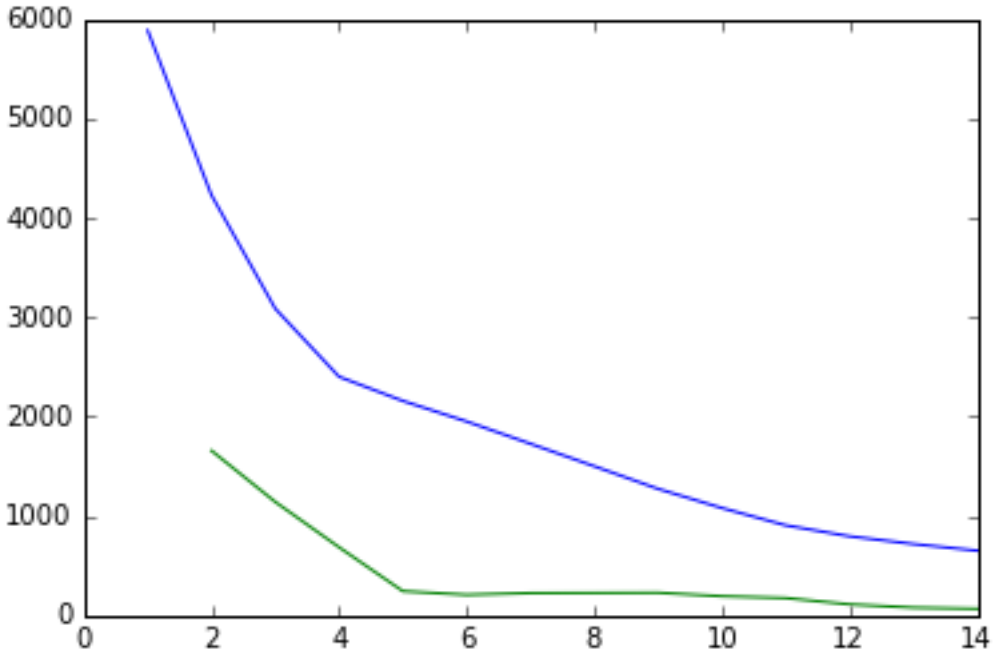
Antes de analizar los datos se procedió a hacer una normalización de éstos y los resultados fueron los siguientes:



Clustering



Gráfica de codo utilizando hierarchy.linkage



Gráfica de codo utilizando algortimo Kmeans

Conclusiones:

Al analizar la forma del dendrograma podemos observar que existen muchos datos con una distancia pequeña entre ellos y grandes saltos de distancia se dan entre cada grupo de datos lo cual indica que se pueden formar grupos de datos adecuadamente con este tipo de agrupamiento.

Con la gráfica de codo y haciendo uso de la gráfica de aceleración se pueden identificar 6 grupos de datos diferentes aunque el algoritmo de kmeans sugiere 5 grupos.

Lo anterior significa que los métodos propuestos para el agrupamiento de datos para la base de datos utilizada logran diferenciar grupos que toman en cuenta tanto el número de trabajadores asegurados dentro de cada industria como el mes y al estar normalizados los datos la distancia ellos no está influenciada por el tamaño de la industria sino por lo que se puede interpretar como la estacionalidad de éstas.