Análisis multivariante aplicado a encuesta nacional de ocupación y empleo (ENOE-INEGI)

Alvarez Sabouret Martín
Facultad de Ciencias Económicas de la Universidad Nacional de Tucumán, Argentina
martinras92@gmail.com

Contenido

INTRODUCCIÓN

La ciencia de datos es hoy en día la herramienta fundamental para la explotación de datos y la generación de conocimiento. Entre los objetivos que persigue se encuentra la búsqueda de modelos que describan patrones y comportamientos a partir de los datos con el fin de tomar decisiones o hacer predicciones. Es un área que ha experimentado un enorme crecimiento al extenderse el acceso a grandes volúmenes de datos e incluso su tratamiento en tiempo real, requiriendo de técnicas sofisticadas que puedan tratar con los problemas prácticos como escalabilidad, robustez ante errores, adaptabilidad con modelos dinámicos.
Ante el acceso a la encuesta nacional de ocupación y empleo del instituto INEGI, me surgió el interés de poder analizar la misma y el hecho de poder emplear herramientas de análisis multivariante y el uso de R-Studio aprendidos a lo largo del cursado de Análisis Cuantitativo II.

DESARROLLO DEL PROBLEMA

Ante la obtención de una base de datos de acceso público, se presenta el desafío poder trabajar sobre la misma empleando técnicas como ser análisis de correspondencias y correlación múltiples, como también estadísticas descriptivas mediante el empleo de software estadístico RStudio.

PREGUNTAS DE INVESTIGACIÓN

¿Existen correspondencias entre las variables categóricas analizadas?

¿Cuáles son aquellas variables más explicativas en el fenómeno de ingresos salariales?

¿Existen diferencias en función del sexo de los individuos?

¿Acaso se cumple la premisa de que, “a mayores años de estudio mejores ingresos”?

OBJETIVO GENERAL

El objetivo es obtener información estadística sobre las características de ocupación y empleo de la muestra bajo estudio, así como información sociodemográfica y económica, con el fin de analizar la estructura laboral y ocupacional de la misma

OBJETIVOS ESPECÍFICOS

Aplicar estadística descriptiva, correspondencias múltiples, clústers, regresión lineal múltiple en busca de realizar un análisis más exhaustivo.

Lograr captar información sobre el tamaño, composición y distribución de la fuerza de trabajo y de los niveles de participación en la actividad económica, sus características y de los niveles de ingresos, estudios y cargo y nivel de salarios de las personas.

MARCO METODOLÓGICO

El trabajo se lleva a cabo empleando una investigación del tipo cuantitativo, no experimental de corte transversal dado que se emplea una muestra sobre un trimestre específico.

Se toma la base de datos ENOE 2021 (encuesta nacional de ocupación y empleo) perteneciente al INEGI (Instituto nacional de estadística y geografía, México), la cual cuenta con más de 10mil registros con 11 variables que se desarrollaran más adelante.

Se pretende llevara a cabo un tipo de investigación exploratorio y explicativo con la intención de abordar la situación laboral de los sujetos encuestados en el censo en busca de patrones y relaciones entre las variables del DataFrame.

Con respecto a las técnicas de análisis, se emplearán estadística descriptiva, análisis multivariante como ser el análisis de correspondencia múltiple, clústers, regresión lineal múltiple con la implementación del software estadístico R.

MARCO TEÓRICO

Análisis multivariante (Hair, Anderson, Tatham, & Black, 1999)

Gran parte de esta creciente comprensión y pericia en el análisis de datos ha venido a través del estudio de la estadística y de la inferencia estadística.
Igualmente, importante, sin embargo, ha sido el dilatado conocimiento y aplicación de un grupo de técnicas estadísticas conocidas como análisis multivariante.
Las técnicas del análisis multivariante están siendo ampliamente aplicadas a la industria,
administración y centros de investigación de ámbito universitario. Por otra parte, pocos campos de investigación o estudio han fracasado en integrar las técnicas multivariantes en su «caja de herramientas» analítica.
los directivos de empresa o los funcionarios de la administración pública, sea cual sea su entorno, que tienen que desarrollar sus conocimientos del análisis multivariante para conseguir una mejor comprensión de los complejos fenómenos de sus ámbitos de trabajo.
Cualquier investigador que examina sólo relaciones de dos variables y que evita el análisis multivariante está ignorando poderosas herramientas que podrían suministrar información potencialmente útil.
Esos métodos hacen posible plantear preguntas específicas y precisas de considerable complejidad en marcos idóneos, lo que posibilita llevar a cabo investigaciones teóricamente significativas y evaluar los efectos de las variaciones paramétricas ocurridas de forma natural en el contexto en que normalmente ocurren.
De esta forma, se pueden preservar las correlaciones naturales entre las múltiples influencias sobre el comportamiento y se pueden estudiar estadísticamente los efectos aislados de esas influencias sin provocar el típico aislamiento de esos individuos o variables.
En un sentido amplio, se refiere a todos los métodos estadísticos que analizan simultáneamente medidas múltiples de cada individuo u objeto sometido a

investigación. Cualquier análisis simultáneo de más de dos variables puede ser considerado aproximadamente como un análisis multivariante. En sentido estricto, muchas técnicas multivariantes son extensiones del análisis univariante (análisis de distribuciones de una sola variable) y del análisis bivariante (clasificaciones cruzadas, correlación, análisis de la varianza y regresiones simples utilizadas para analizar dos variables).

Análisis de correspondencias (Hair, Anderson, Tatham, & Black, 1999)

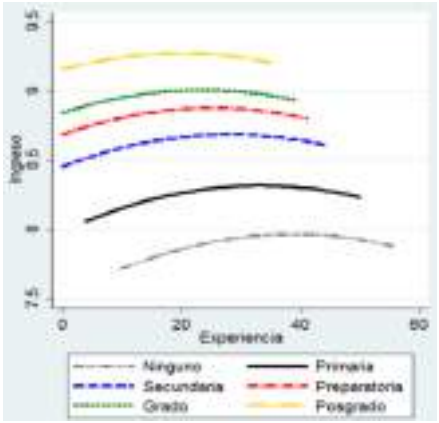
El análisis de correspondencias es una técnica de interdependencia recientemente desarrollada que facilita tanto la reducción dimensional de una clasificación de objetos (por ejemplo, productos, personas, etc.,) sobre un conjunto de atributos y el mapa perceptual de objetos relativos a estos atributos. Los investigadores se enfrentan constantemente a la necesidad de «cuantificar datos cualitativos» que encuentran en variables nominales. El análisis de correspondencias difiere de otras técnicas de interdependencia discutidas antes en su capacidad para acomodar tanto datos no métricos como relaciones no lineales.
En su forma más básica, el análisis de correspondencias emplea una tabla de contingencia.

Ecuación de Mincer (Mincer, Jacob,1958)

La ecuación de Mincer es, en economía del trabajo y en economía de la educación, una relación matemática que relaciona el número de años de estudio y el número de años de experiencia con el salario de un individuo en el mercado laboral. Trabajo, nombrado en memoria de Jacob Mincer. Se utiliza para medir la compensación que hacen los agentes entre la búsqueda de una educación superior y la entrada directa al mercado laboral. La ecuación se puede notar con ingresos, ingresos para una persona sin educación o experiencia, el número de años de escolaridad y el número de años de experiencia.

$$\ln W = \ln W_0 + \rho S + \beta_1 X + \beta_2 X^2$$

Dónde las variables tienen los significados siguientes W son los ingresos Wo son los ingresos de alguien sin ninguna educación y ni experiencia); S son los años de escolarización; X son los años de experiencia laboral potencial en el mercado. Los parámetros ρ, β1, β2 pueden ser interpretados como los retornos a escolarización y experiencia, respectivamente.



Fuente: "Función_de_ingreso_de_Minceer (wikimedia.org)"

Análisis clúster (Hair, Anderson, Tatham, & Black, 1999)

El análisis clúster es una técnica analítica para desarrollar subgrupos significativos de individuos u objetos. De forma específica, el objetivo es clasificar una muestra de entidades (personas u objetos) en un número pequeño de grupos mutuamente excluyentes basados en similitudes entre las entidades. En el análisis clúster, a diferencia del análisis discriminante, los grupos no están predefinidos.
Por consiguiente, se usa la técnica para identificar los grupos.
Habitualmente, el análisis clúster implica al menos dos etapas. La primera es la medida de alguna forma de similitud o asociación entre las entidades para determinar cuántos grupos existen en realidad en la muestra. La segunda etapa es describir las personas o variables para determinar su composición. Este paso puede llevarse a cabo aplicando el análisis discriminante a los grupos identificados por la técnica clúster.
que es la tabulación cruzada de dos variables categóricas. A continuación, transforma los datos no métricos en un nivel métrico y realiza una reducción dimensional (similar al análisis factorial) y un mapa perceptual (similar al análisis multidimensional).

Clustering (García, Molina, & Bustamante, 2018)

También llamadas *técnicas de agrupamiento* o *técnicas de segmentación*, permiten la identificación de tipologías o grupos donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos. Así se puede segmentar el colectivo de clientes, el conjunto de valores e índices financieros, el espectro de observaciones astronómicas, el con junto de zonas forestales, el conjunto de empleados y de sucursales u oficinas, etc. La segmentación está teniendo mucho interés desde hace ya tiempo dadas las importantes ventajas que aporta al permitir el tratamiento de grandes colectivos de forma pseudo particularizada, en el más idóneo punto de equilibrio entre el tratamiento individualizado y aquel totalmente masificado.
Las herramientas de *clustering* pueden estar basadas en múltiples tipos de técnicas, como estadística, empleo de algoritmos matemáticos o generación de reglas o de redes neuronales para la trata miento de registros. Para otro tipo de elementos a agrupar o segmentar, como texto y documentos, se usan técnicas de reconocimiento de conceptos.
Esta técnica suele servir de punto de partida para después hacer un análisis de clasificación sobre los clústeres.
La principal característica de esta familia de técnicas es la utilización de una medida de similaridad que, en general, está basada en los atributos que describen a los objetos, y se define usualmente por proximidad en un espacio multidimensional. Para datos numéricos, suele ser preciso preparar los datos antes de realizar *datamining* sobre ellos, de manera que en primer lugar se someten a un proceso de estandarización.

Regresión múltiple (Hair, Anderson, Tatham, & Black, 1999)

La regresión múltiple es el método de análisis apropiado cuando el problema del investigador incluye una única variable métrica dependiente que se supone está relacionada con una o más variables métricas independientes. El objetivo del análisis de la regresión múltiple es predecir los cambios en la variable dependiente en respuesta a cambios en varias de las variables independientes.

Este objetivo se consigue muy a menudo a través de la regla estadística de los mínimos cuadrados.
La regresión múltiple es útil siempre que el investigador esté interesado en predecir la cantidad o la magnitud de la variable dependiente. Por ejemplo, se puede hacer la predicción de los gastos mensuales de cenar fuera de casa (variables dependientes) con información referente a la renta familiar, su tamaño y la edad del cabeza de familia (variables independientes). De la misma forma, el investigador puede intentar predecir las ventas de una compañía a partir de información sobre sus gastos en publicidad, el número de vendedores y el número de tiendas que distribuyen sus productos.

DESARROLLO

El trabajo ser realizó sobre la base de datos proporcionada de manera pública por (INEGI) Instituto Nacional de Estadística y Geografía, es uno de los órganos constitucionales autónomos de México con gestión, personalidad jurídica y patrimonio propios, responsable de normar y coordinar el Sistema Nacional de Información Estadística y Geografía.

Se encarga de realizar los censos nacionales; integrar el sistema de cuentas nacionales y estatales (es decir, el flujo de producción, consumo y distribución de la actividad económica); y, desde 2011, elabora los índices nacionales de Precios al Consumidor, e Índice Nacional de Precios al Productor.

Específicamente se trabajó sobre la sección de Encuesta Nacional de Ocupación y Empleo (ENOE), la cual busca medir el salario mensual (ingreso mensual) de los mexicanos y es. Esta encuesta se realiza cada trimestre. Se tomó una muestra de la encuesta realizada en el cuarto trimestre de 2021.

Aclaraciones sobre las variables

A la hora de definir la variable sexo, el instituto informa lo siguiente: La cuestión de si en las estadísticas de hombres y mujeres se debe referir a “género” o a “sexo” ha quedado esclarecida en los foros internacionales que abordan este tema. Según la oficina de Estadísticas de Suecia, la palabra “sexo” hace referencia a las diferencias biológicas entre hombres y mujeres; mientras que “género” es una construcción social (normas, costumbres y prácticas) y una condición de las diferencias entre los sexos y de las relaciones sociales entre hombres y mujeres, ya que la identidad social de género depende de factores ideológicos, históricos, culturales, religiosos, étnicos y económicos. En México, para las estadísticas sobre establecimientos y empresas, particularmente las de los Censos Económicos, cuando presentan el personal ocupado que trabaja en las unidades económicas, se asume que la desagregación de esa variable en hombres y mujeres, es una distinción de sexo. El eje de este producto son los hombres y las mujeres, desde el punto de vista de su inserción en los puestos de trabajo que ocupan en las unidades económicas. Por lo anterior, es importante señalar que la desagregación de hombres y mujeres que se hace en este documento está en

función de las diversas categorías de personal ocupado total que se distinguen en los Censos Económicos.

El data frame puntualmente posee 10.280 registros con 11 variables las cuales son:

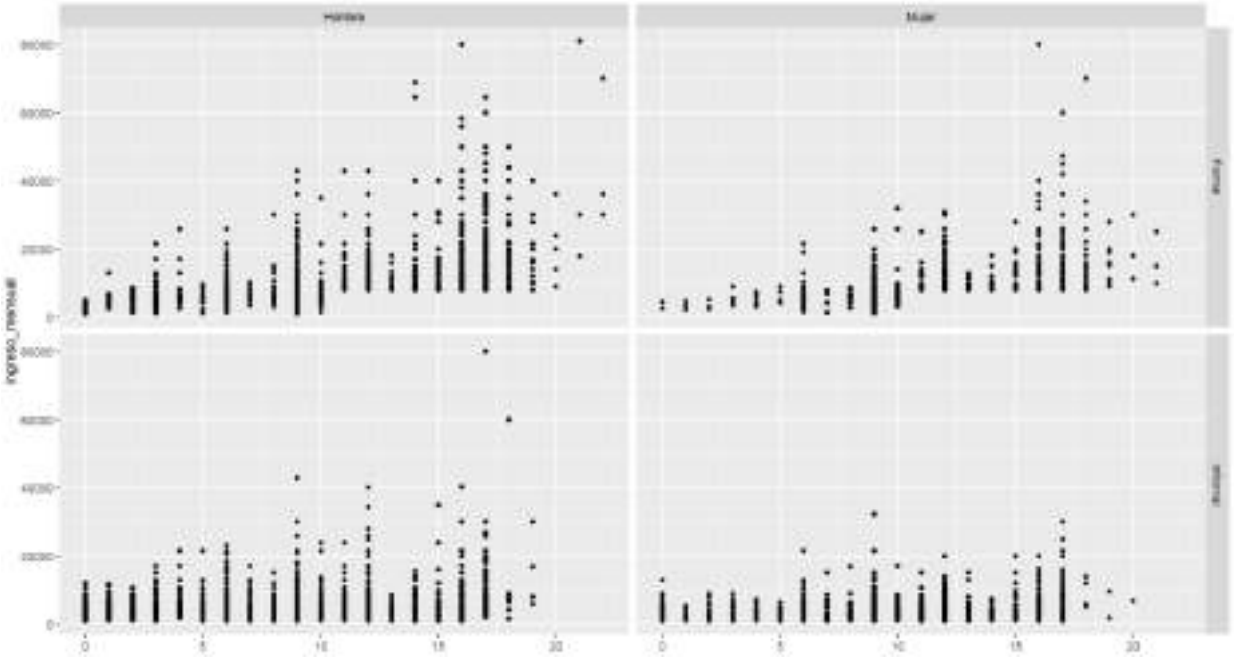
1. **Estado**, el cual indica el área geográfica de residencia del sujeto entrevistado.
2. **Sex**, discriminando por sexo (mujer, hombre).
3. **Edad**.
4. **Pos_ocu**, la misma clasifica al individuo según sea:
    a. “Rabajador subordinado y remunerado”;
    b. “Empleador”; o
    c. “Trabajador por cuenta propia”.
5. **Ing_salarios**: indica el rango en el que se encuentra el salario percibido en comparación con el salario mínimo. Por lo cual se tienen las siguientes categorías:
    a. Hasta un salario mínimo.
    b. Más de 1 hasta 2 salarios mínimos.
    c. Mas de 2 hasta 3 salarios mínimos.
    d. Más de 3 hasta 5 salarios mínimos.
    e. Más de 5 salarios mínimos.
6. **Niv_edu:** indica el nivel educativo:
    a. Primaria incompleta
    b. Primaria completa
    c. Secundaria completa
    d. Medio superior y superior
7. **Anios_esc:** cantidad de años de estudios, en la muestra analizada el intervalo de los sujetos tiene un mínimo de 0años y máximo de 22años.
8. **Hesocup:** la cual indica las horas semanales destinadas al trabajo.
9. **Ingreso_mensual:** el salario percibido en moneda local de forma mensualizada.
10. **Num_trabajos:** indica la cantidad de trabajos que posee la persona siendo el máximo de trabajos de la muestra de dos empleos.
11. **Tipo_empleo:** categorizando por Formal o Informal.

A continuación, se muestra una captura de la base de datos para tener una mejor noción de la misma.

ANÁLISIS DESCRIPTIVO

En el siguiente grafico de dispersión se trabajó con las variables **sexo**, **ingresos**, y **tipo de empleo**. Se puede apreciar las diferencias salariales entre los hombres y las mujeres, donde en términos generales los ingresos de los hombres son más altos sobre todo cuando el tipo de empleo es **Informal**

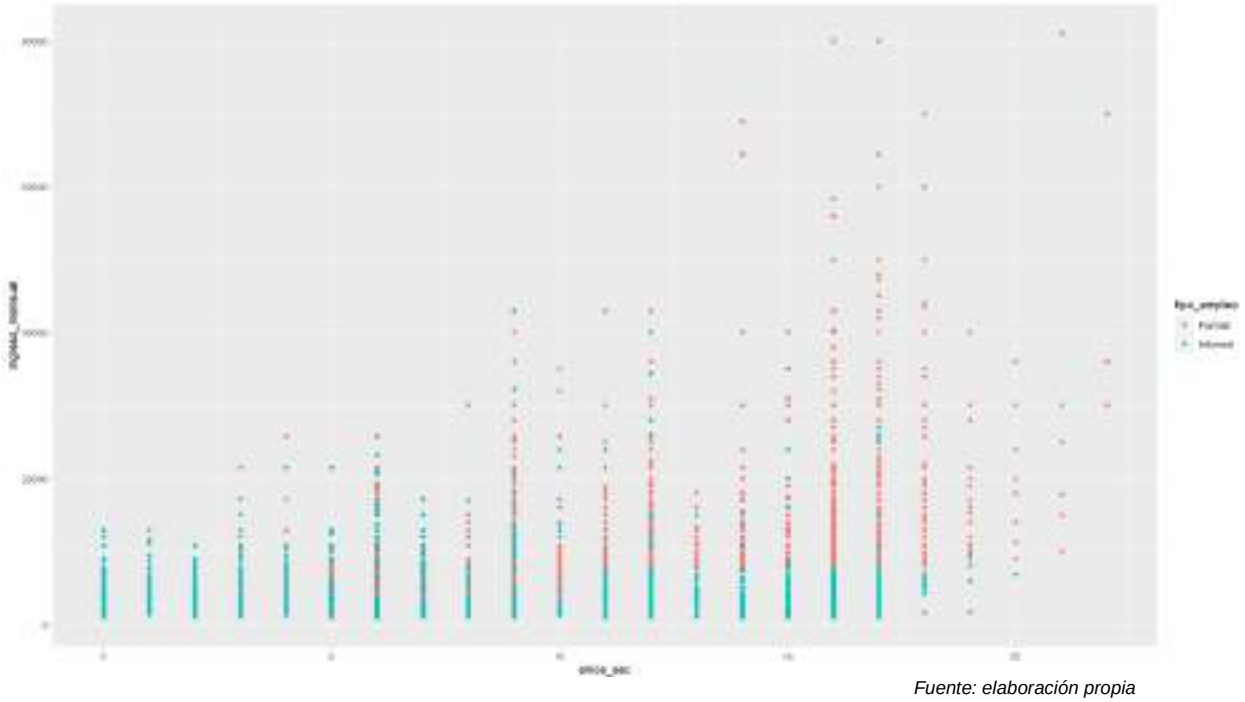
Gráfico de dispersión matricial empleando variables sexo, ingresos mensuales, tipo de empleo



Fuente: elaboración propia

Al realizar un gráfico de dispersión empleando las variables años de estudios, tipo de empleo e ingresos mensuales, se aprecia una correlación positiva entre los años de estudio y los ingresos y una relación negativa entre tipo de empleo “informal” y años de estudio, es decir, a mayores años de estudio se aprecia mayores ingresos, a su vez a mayores años de estudio el tipo de empleo informal va decreciendo tomando más preponderancia el empleo formal. El fenómeno se corresponde tal como lo predicen las ecuaciones de Mincer.

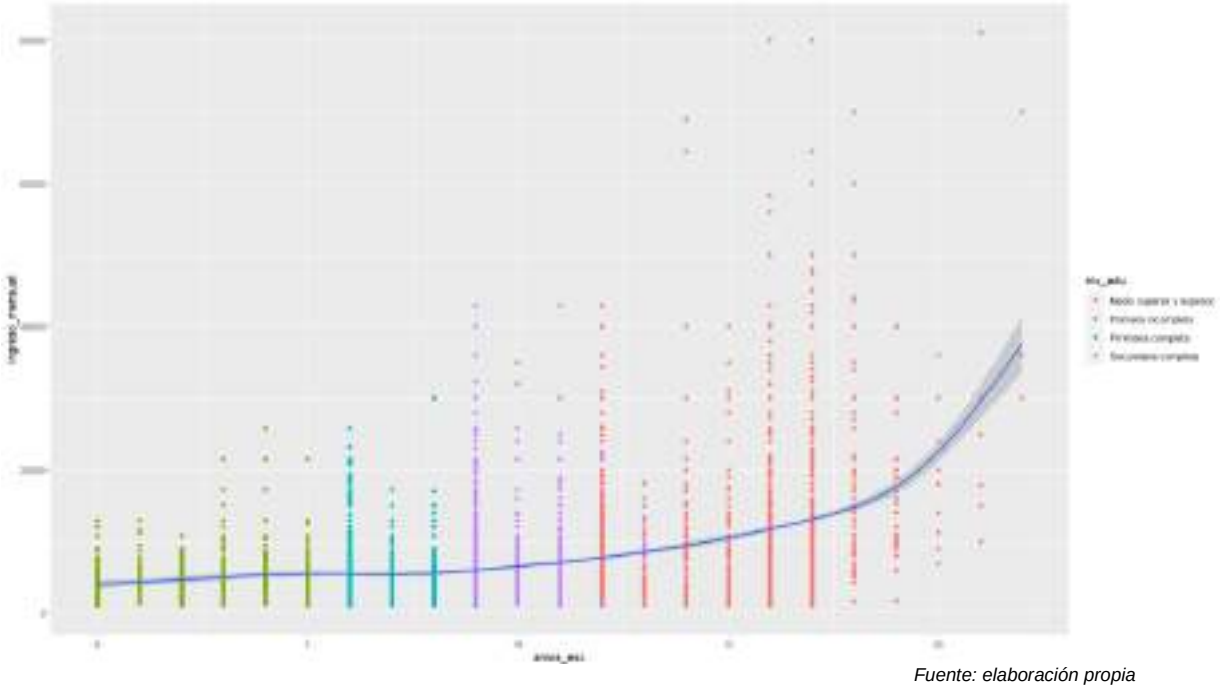
Gráfico de dispersión empleando ingresos mensuales, años de escolaridad, tipo de empleo.



Fuente: elaboración propia

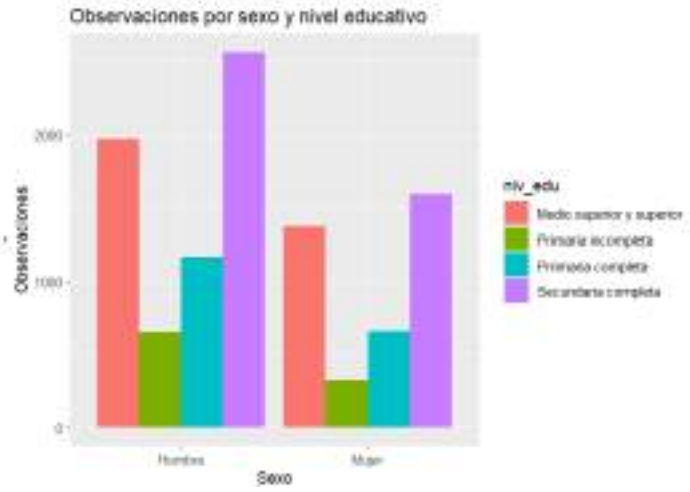
Si incorporamos la categorización en base al nivel educativo y una línea de tendencia, se evidencia la correlación entre los años de estudio y el nivel educativo y la tendencia positiva a poseer mayores ingresos mensuales donde la curva comienza a manifestarse exponencialmente a partir de los 12 años de estudio.

Gráfico de dispersión con línea de tendencia empleando las variables Ingresos mensuales, Nivel educativo, años de educación.



Fuente: elaboración propia

En la siguiente gráfica se representa el nivel de estudio discriminado en función del sexo

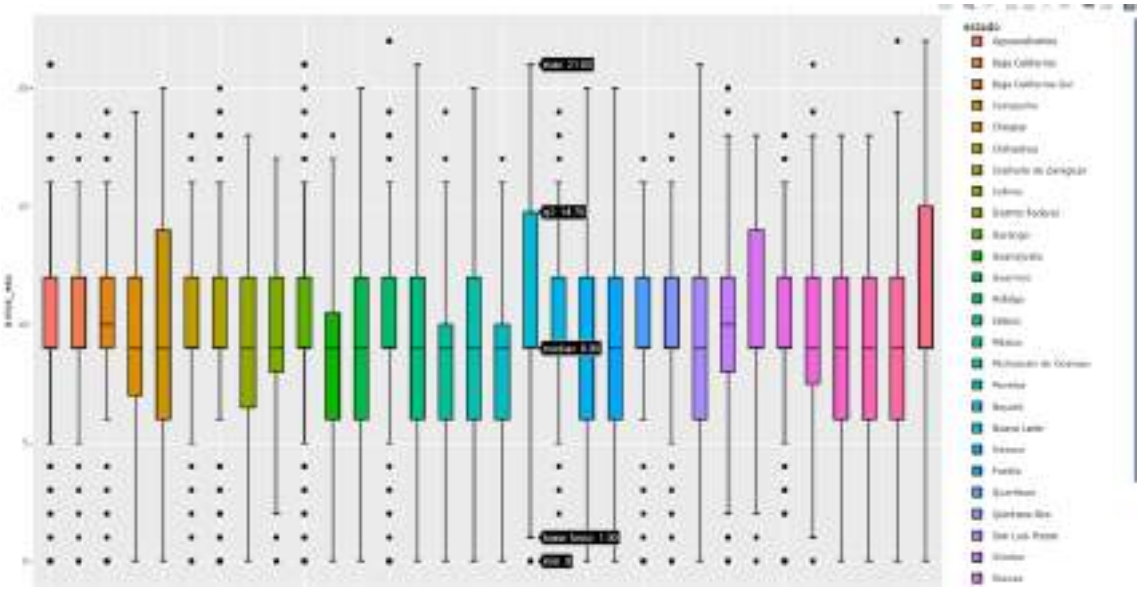


Fuente: elaboración propia

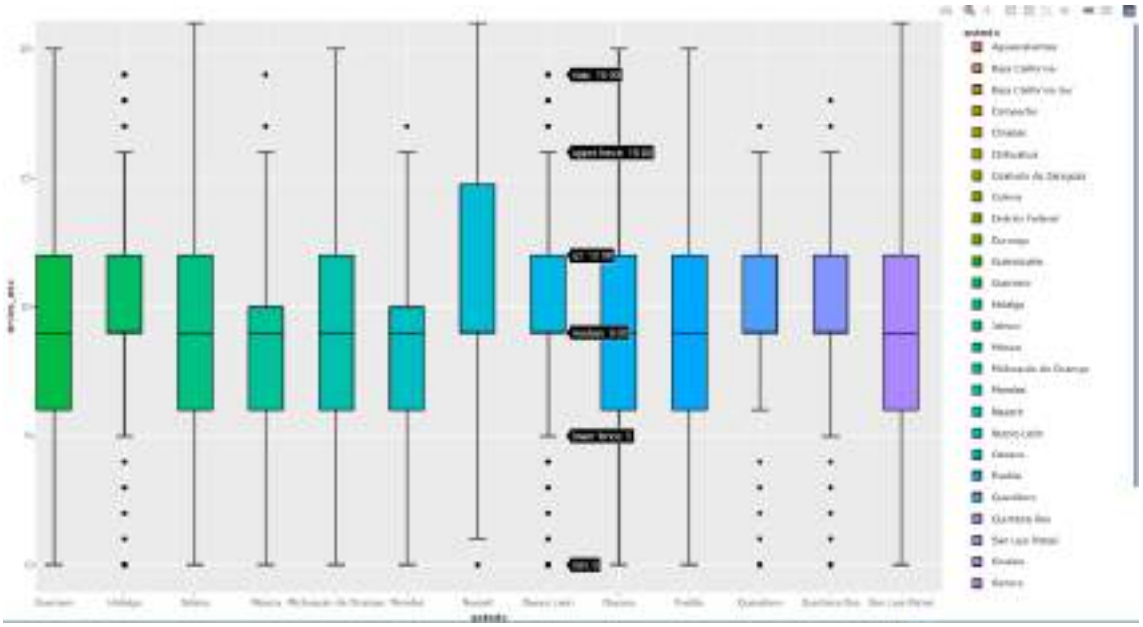
Con la intención de visualizar cuáles son aquellos estados que poseen mayor cantidadad de años de estudio, se empleo la función “ggploty” que permite visualizar

claramente dentro de los diagramas de caja representados la información pertinente a la mediana, los cuartiles, los máximos y mínimos y si existiesen outliers.

BOXPLOTS POR ESTADO



Fuente: elaboración propia



Fuente: elaboración propia

Analizando los estados, se aprecia que dentro de los que poseen mayores años de estudio lo encabezan de mayor a menor cantidad de años: la Ciudad de México, Nuevo León y Sonora.

Mientras que los estados que presentan la menor cantidad de años de estudios de sus habitantes, ordenados de menor a mayor son los siguientes:

Chiapas, seguido de Oaxaca, Guerrero, Michoacán, Veracruz, Guanajuato, Zacatecas, Hidalgo, Yucatán y San Luis Potosí.

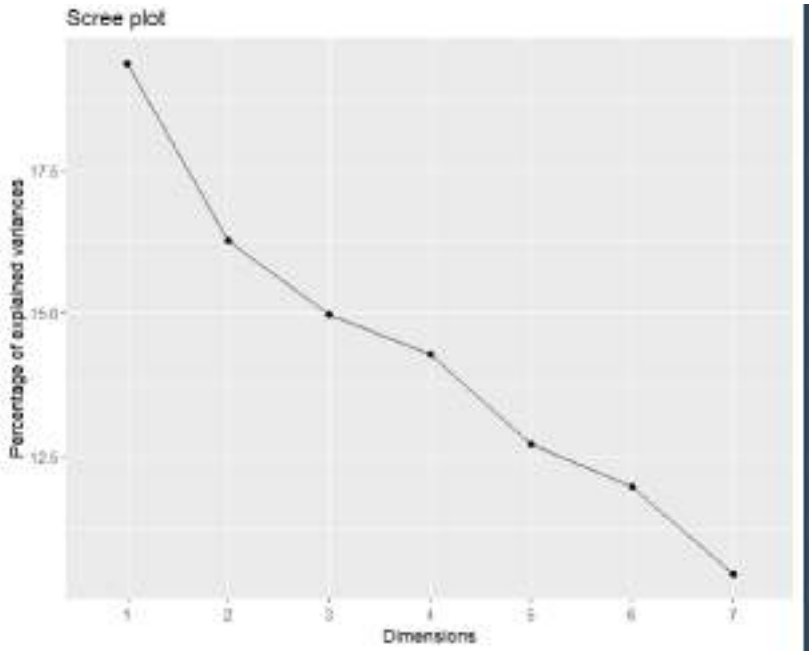
APLICACIÓN DE ANÁLISIS DE CORRESPONDÉNCIAS MULTIPLE (MCA)

MCA empleando las variables sexo, posición ocupada y categoría de ingresos

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
sex	0.3685904	0.003683828	0.1475355	2.458417e-31	0.1845988
pos_ocu	0.4103937	0.577227422	0.1228955	1.561746e-25	0.2998977
ing_salarios	0.5750452	0.557411953	0.7770748	1.000000e+00	0.4057946

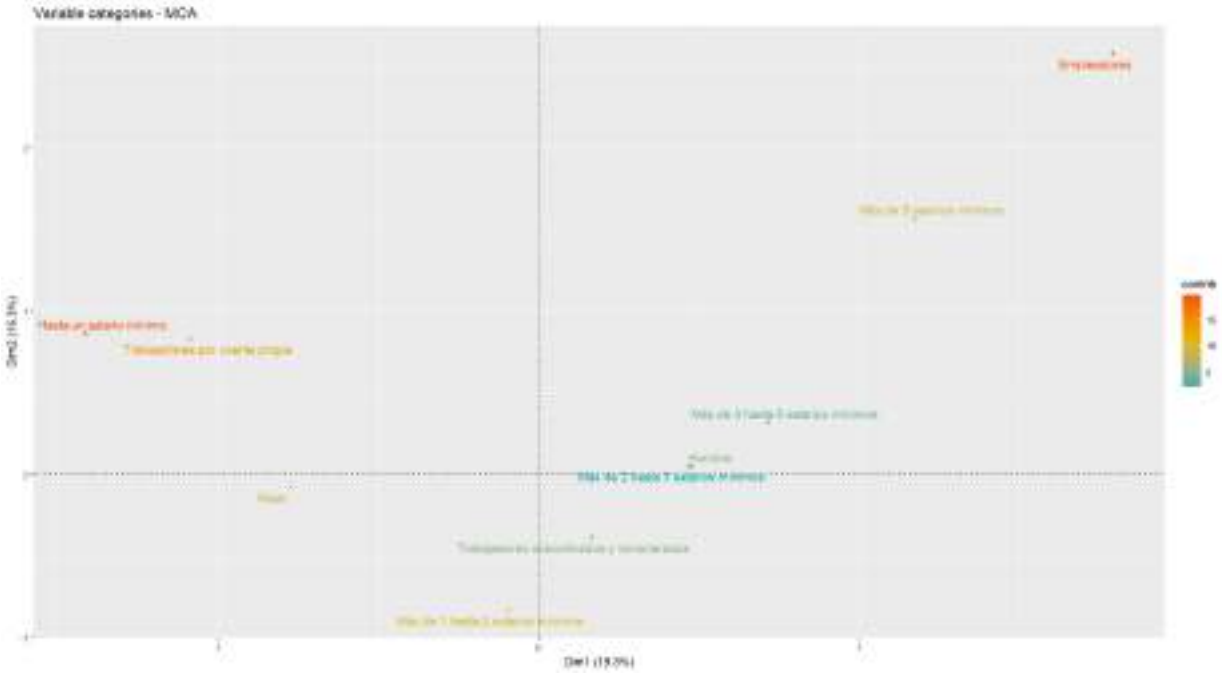
Fuente: elaboración propia

Con el empleo de las variables mencionadas, se procede a el cálculo de la significancia que aportan las dimensiones generadas. A continuación, se aprecia el porcentaje de explicación de las variaciones, Con las dimensiones 1 y 2 se acumula una explicación del 40% aproximadamente.



Fuente: elaboración propia

Figura: MCA1



Fuente: elaboración propia

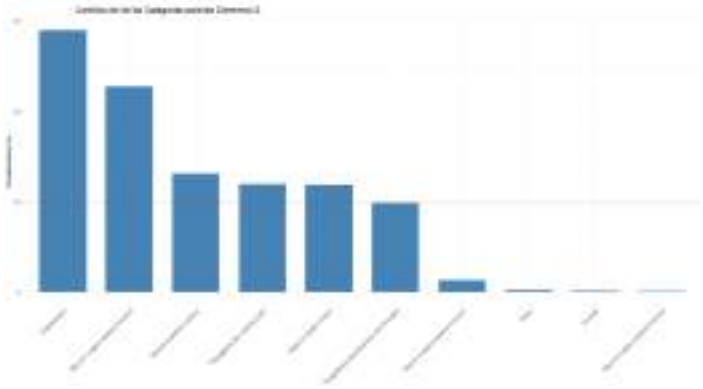
Observaciones de la figura MCA1, recordemos que las variables en juego son el sexo, la ocupación en el empleo y el rango de ingreso salarial mensual. Las dimensiones bajo estudio explican alrededor del 40% del fenómeno.

Podemos apreciar la correspondencia que existe entre los individuos que son empleadores quienes poseen mayor tendencia a tener un rango de ingresos superior al de 5 salarios mínimos.

Al otro extremo se encuentran aquellos trabajadores por cuenta propia que se corresponden con los ingresos más bajos, es decir, “hasta un salario mínimo”.

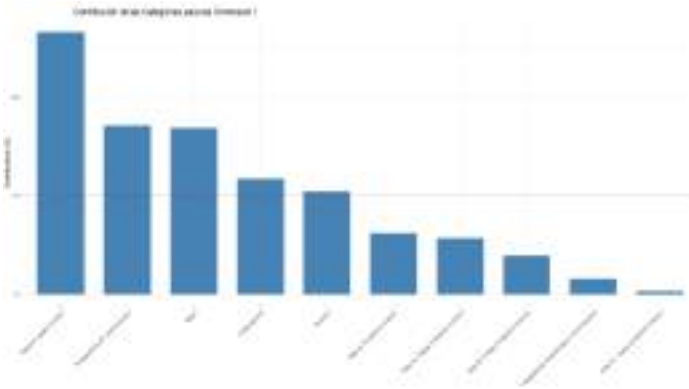
Si bien cuesta realizar otras correspondencias dado que las demás variables intermedias en base al mapa de calor, no tienen grandes tasas de contribución a las dimensiones. Sin embargo, observamos que los hombres se encuentran más cercanos a percibir más de dos salarios mínimos con mayor proximidad a la posición ocupada en el empleo “subordinados y remunerados”. Mientras que las mujeres se encuentran en el punto medio entre la posición ocupada como trabajadores por cuenta propia y subordinados remunerados y el rango de ingresos entre 1 hasta 3 salario mínimos.

Contribuciones de las variables a la dimensión 2



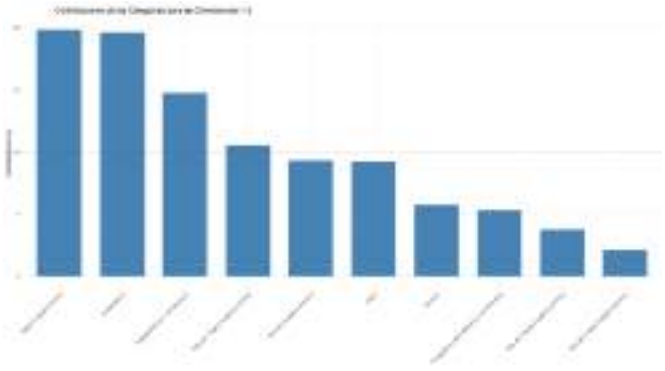
Fuente: elaboración propia

Contribución de las variables a la Dimensión 1



Fuente: elaboración propia

Contribución de las variables a ambas dimensiones

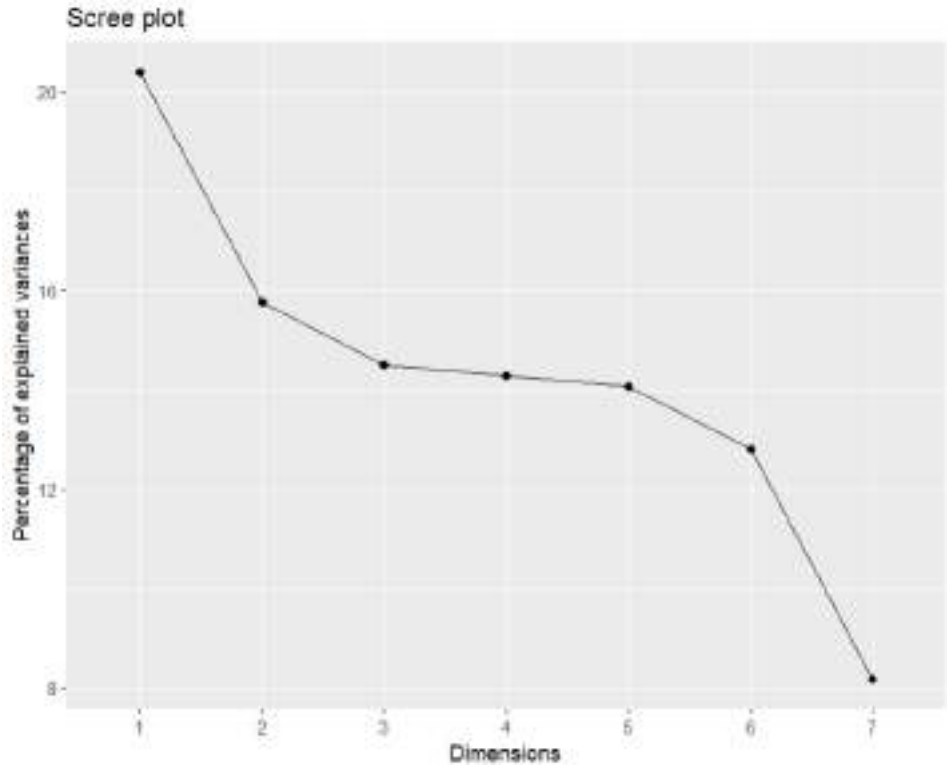


Fuente: elaboración propia

MCA empleando variables ingresos y nivel educativo

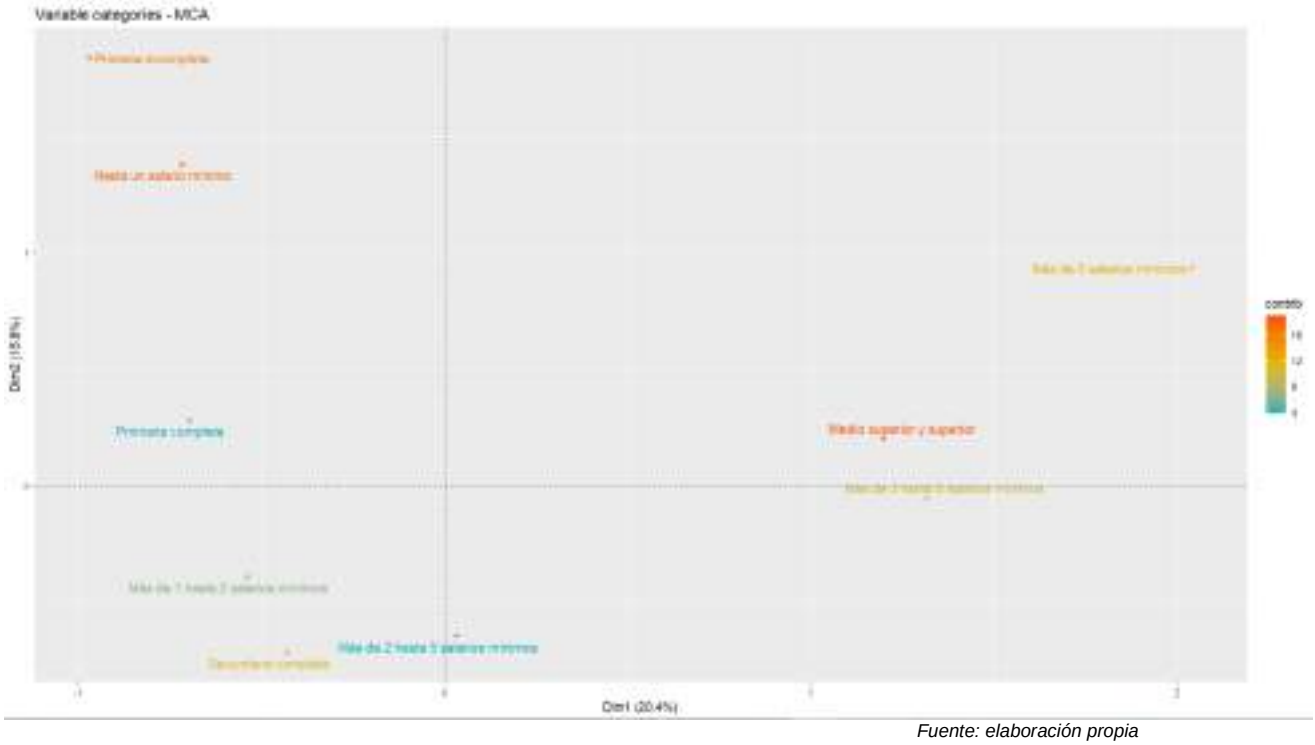
	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
ing_salarios	0.7137726	0.5520722	0.5074478	1.000000e+00	0.4925522
niv_edu	0.7137726	0.5520722	0.5074478	2.906889e-24	0.4925522

Fuente: elaboración propia

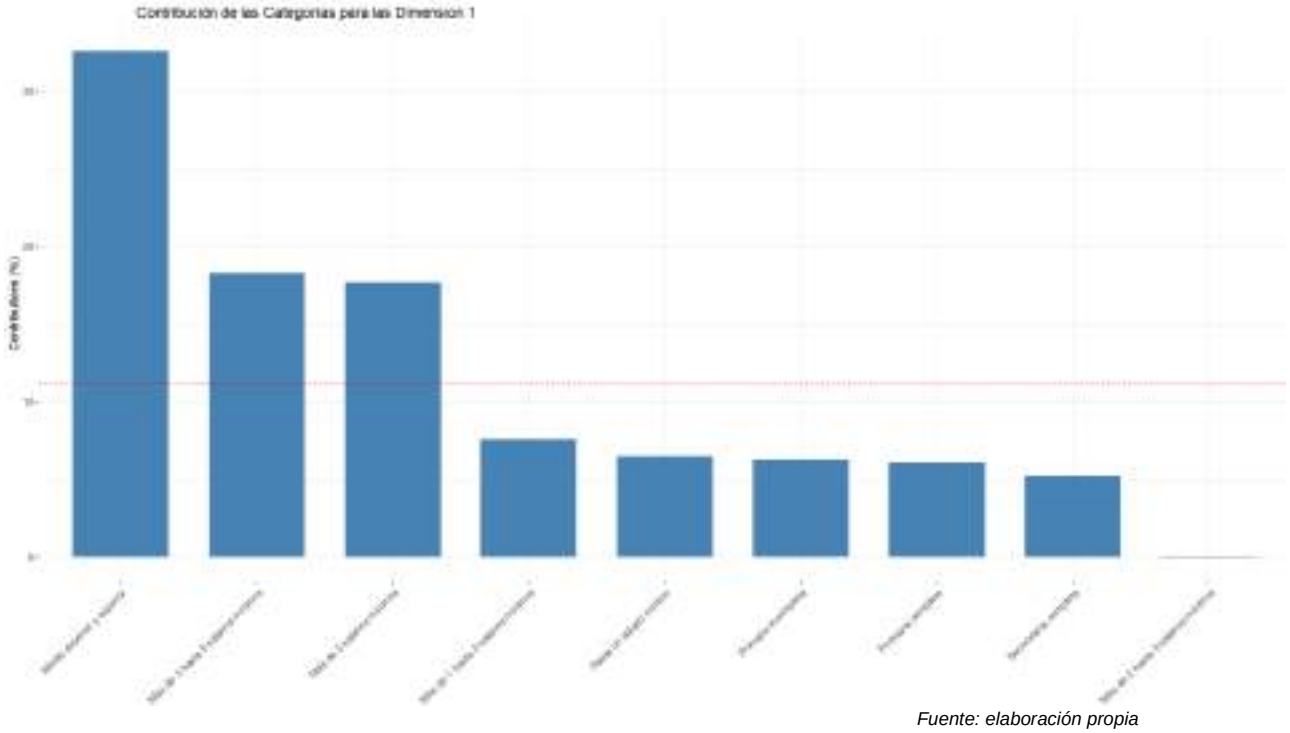


Fuente: elaboración propia

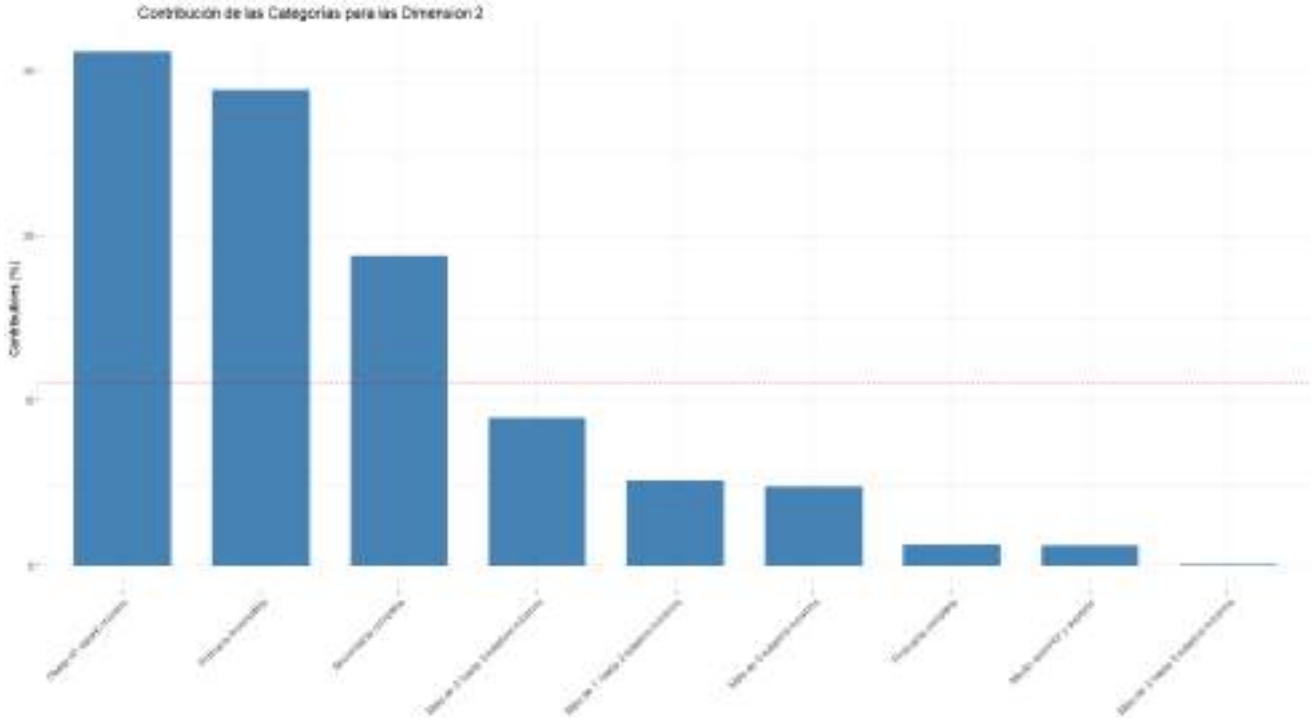
En esta ocasión se analizan las variables ingresos con el nivel educativo, vemos como se corresponde con los estudios de Mincer al respecto del nivel de estudios con el salario esperado, apreciándose una correspondencia clara entre el nivel de estudio medio-superior y superior y los ingresos de 3 a 5 salarios mínimos y superiores a 5 salarios mínimos.



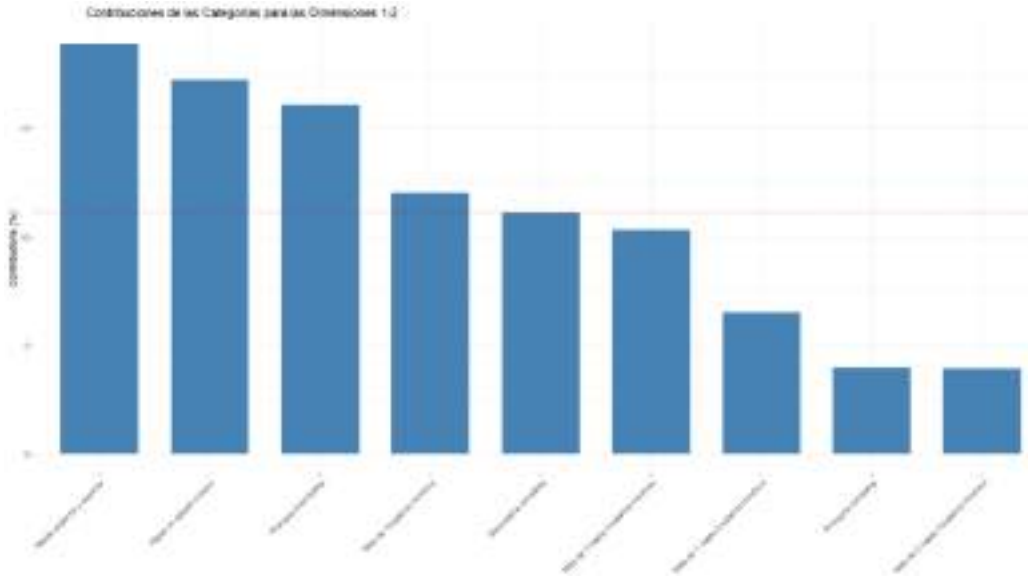
Fuente: elaboración propia



Fuente: elaboración propia



Fuente: elaboración propia

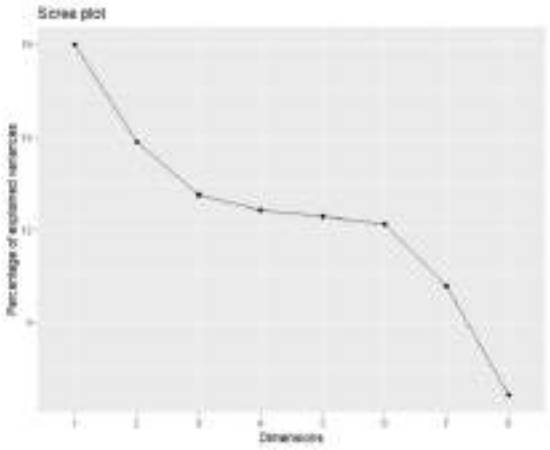


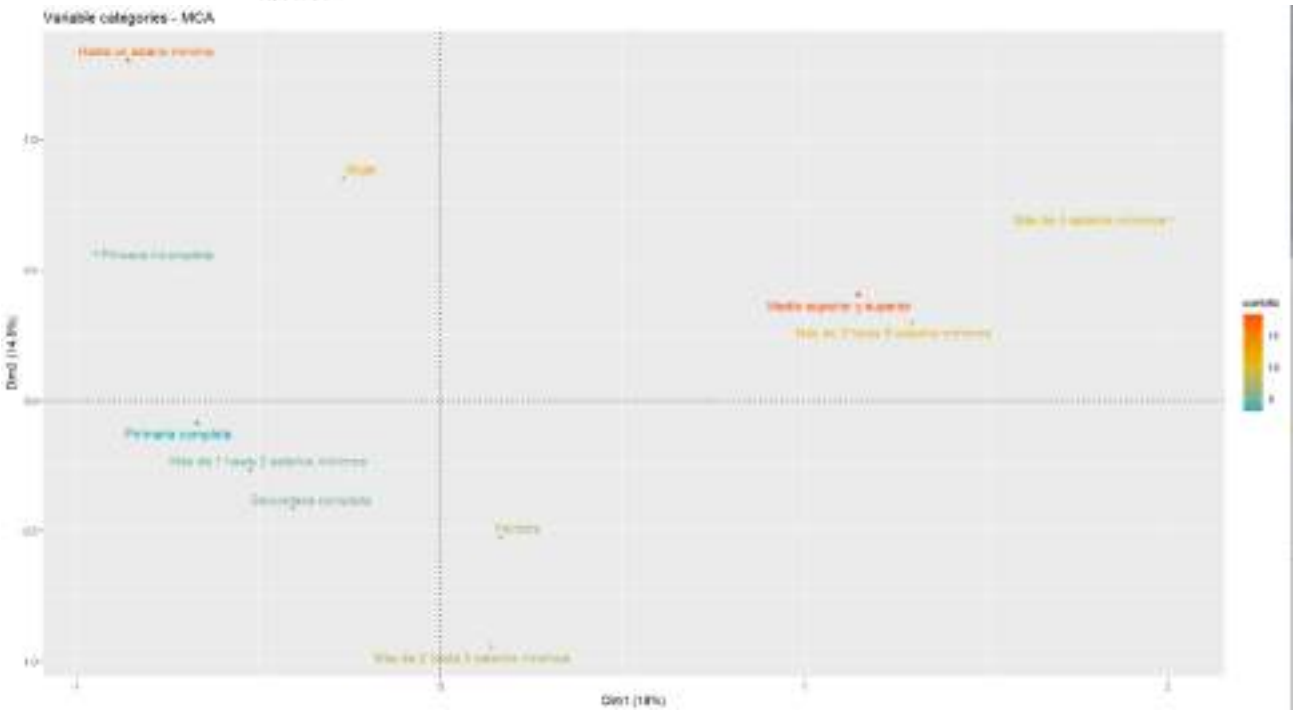
Fuente: elaboración propia

Si procedemos a agregar la variable sexo:

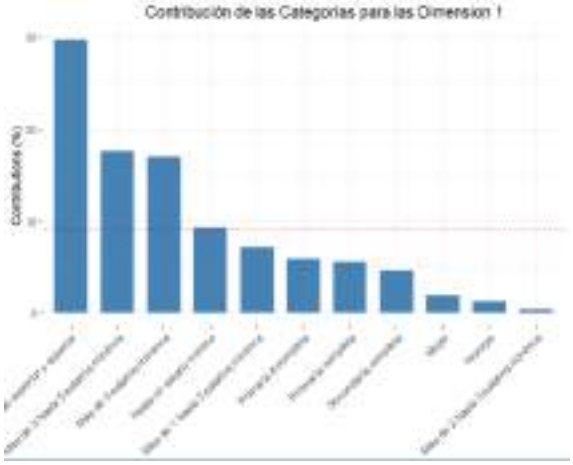
MCA con variables SEXO INGRESOS y NIVEL EDUCATIVO

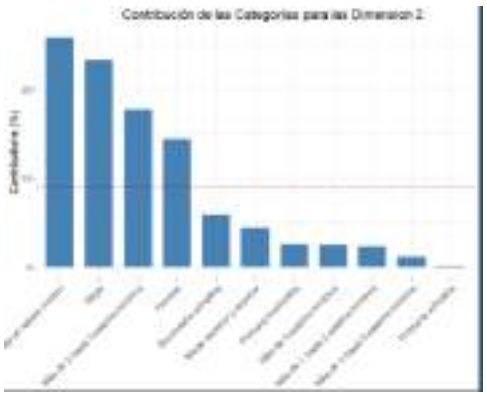
	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
sex	0.04401985	0.4472590	0.1633222	0.01442663	0.0178647
ing_salarios	0.73686660	0.5853750	0.3405291	0.43235279	0.6391154
niv_edu	0.65800100	0.1537156	0.5461953	0.56329740	0.3365933

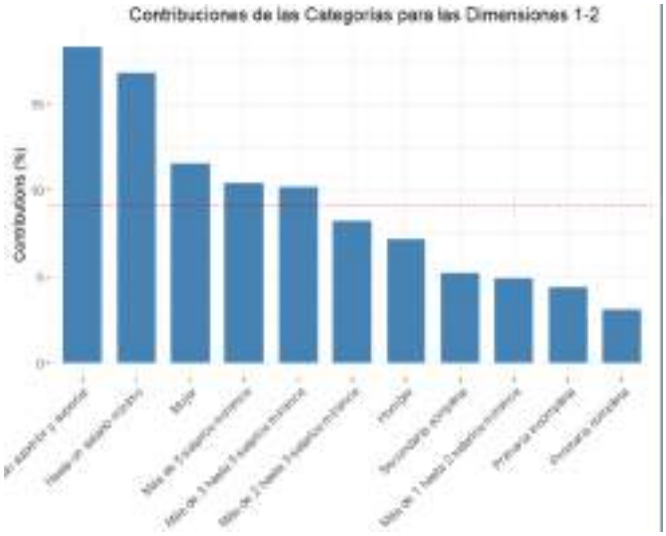




Esta vez con la variable sexo junto a nivel de ingresos y nivel educativo, se aprecia que en la muestra analizada de la población, el sexo femenino se encuentra en el segundo cuadrante junto a salarios bajos y niveles bajos de educación, en cuanto al sexo masculino, se aprecia que se encuentra posicionado en un nivel intermedio mas cercano al nivel educativo “secundaria completa” y con salarios cercanoas a la categoría superior a 2 salarios mínimos.


Contribución de las Categorías para las Dimensión 1
Contributions (%)

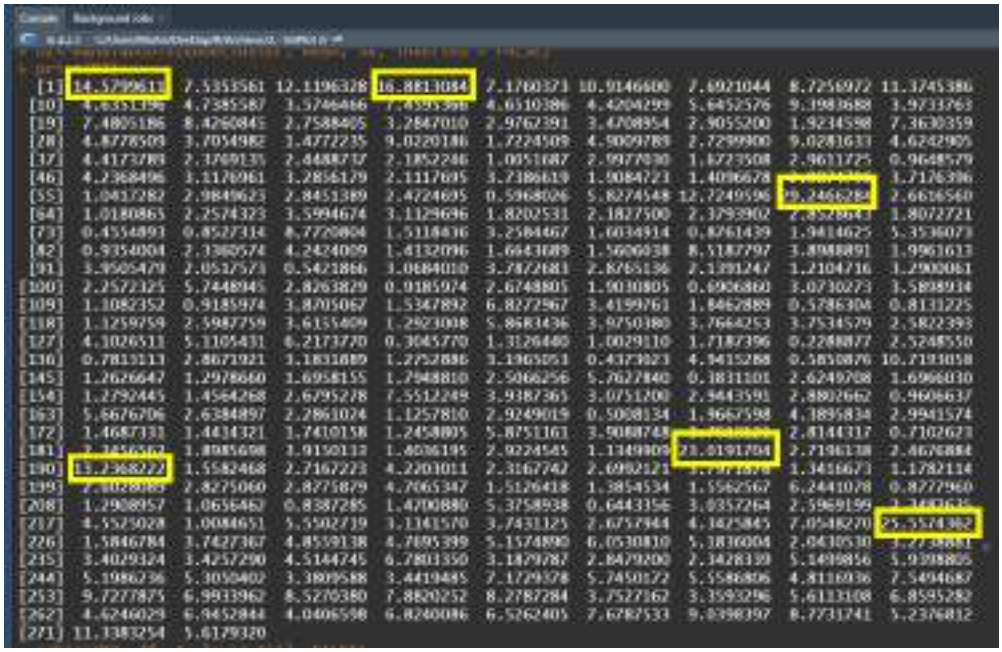

Contribución de las Categorías para las Dimensión 2


Contribuciones de las Categorías para las Dimensiones 1-2
Contributions (%)

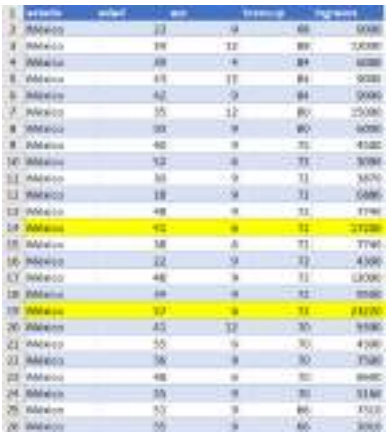
CLUSTER

Para el empleo de clústers se procedió a analizar los estados extremos en cuanto a los años de educación obtenidos previamente en los análisis descriptivos. Es decir, analizaremos el estado de México siendo este el que poseía mayores años de estudio y el estado de Sonora que contaba con los menores años de estudio.

En primera instancia se realizó la limpieza de aquellos datos atípicos que en este caso en base a las distancias de Mahalanobis se determinaron los atípicos como aquellos que superaban el valor de 13,27.



Luego de la detección de 6 outliers, se procedió a eliminar los mismos.



En cuanto a la matriz de distancias se eligieron los números de cluster teniendo en cuenta los indicadores óptimos de numero de clusters con método de distancia euclídea y empleando en el scrip: wss, gap_stat, y silhouette. El mismo correspondía con el número óptimo de dos (2) conglomerados. De esta manera se logra una explicación dimensional del 70%

Clúster elaborado con el estado de México:

Fuente: elaboración propia

Se conformaron 2 grupos

Media de Cada Variable por Conglomerado, Utilizando los datos originales

cluster	edad	anios_esc	hrsocup	ingreso_mensual
1	35.48571	14.657143	43.47619	12284.761
2	41.52174	8.088379	42.09836	5585.012

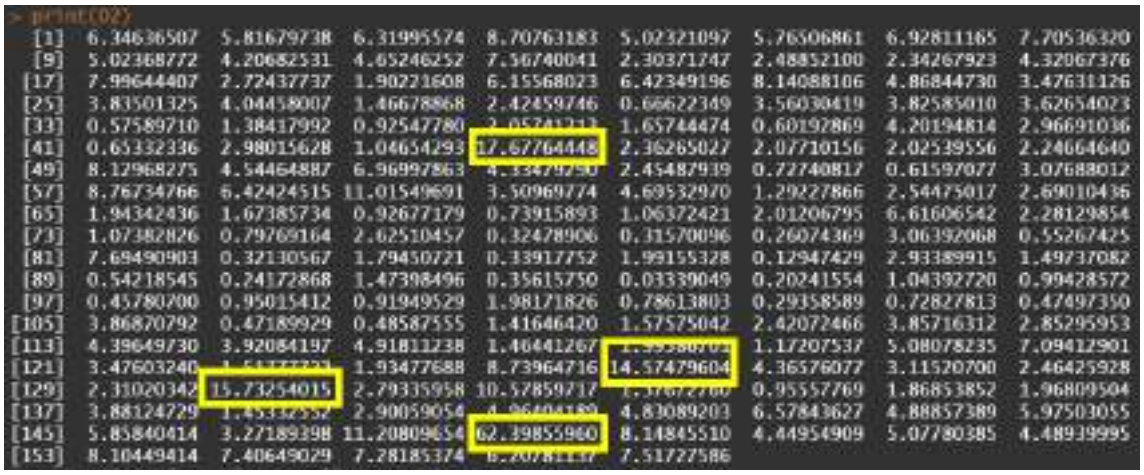
En este caso se aprecia que, de los 2 conglomerados, las variables más representativas son el ingreso junto a los años de estudio, dándose una correlación positiva entre las mismas, a su vez, el primer grupo representado en color rojo del lado izquierdo del gráfico, el análisis de los promedios es, para edad de 35 años, en años de escolaridad 15años, horas de ocupación de 43, y un ingreso alrededor de los $12.285 moneda local.

Mientras que el segundo grupo conformado los promedios son los siguientes, 42 años de edad, 8 años de escolaridad, 42 hs de ocupación laboral, y salarios en promedio de $5585.

Comparación con el estado de Sonora:

El mismo procedimiento empleado en México se llevó a cabo con el estado Sonora a fin de apreciar los contrastes posibles dado que ambos representan los máximos y mínimos en el ranking de los años de estudio respectivamente.

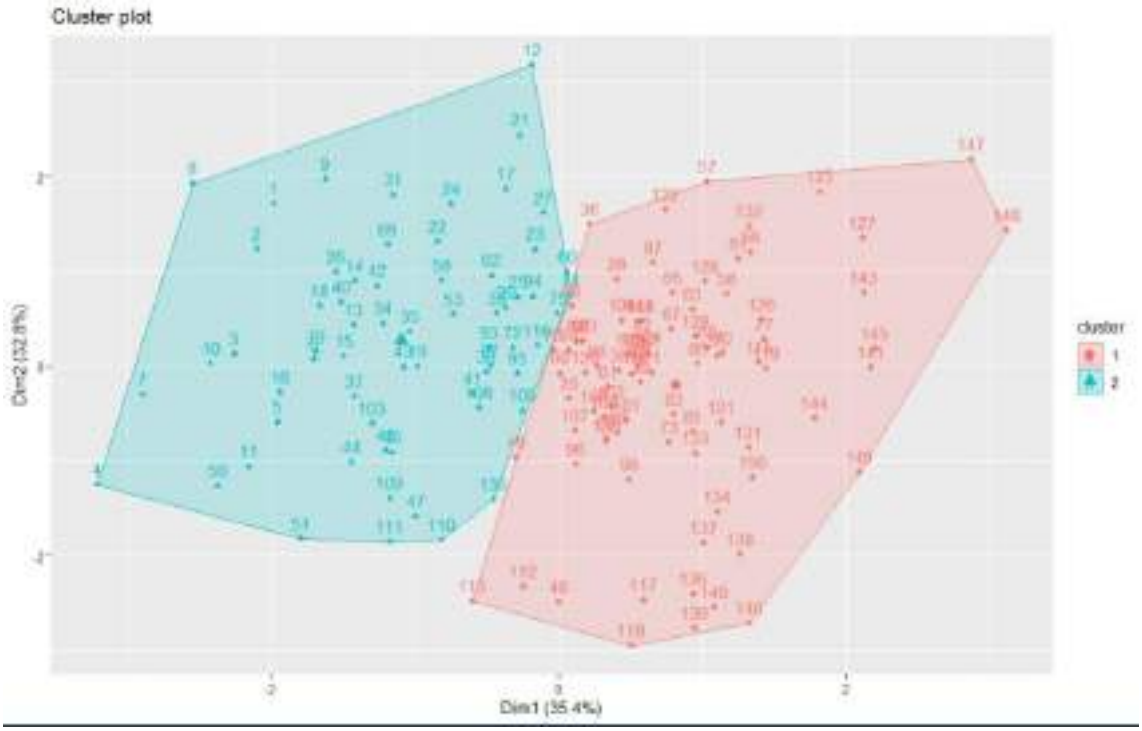
Por lo que, siguiendo el mismo proceso mencionado anteriormente, la matriz de distancia de Mahalanobis arrojada con los datos atípicos es la siguiente:



Eliminación de los 4 outliers detectados:

edad	anios_esc	hrsocup	ingreso_mensual
31	17	96	23000
32	12	84	17200
66	6	84	5000
74	9	84	4600
30	12	84	8000
23	17	84	7740
30	12	84	17200
29	16	84	10500

Clúster elaborado con el estado de Sonora:



Media de Cada Variable por Conglomerado, Utilizando los datos originales

cluster	edad	anios_esc	hrsocup	ingreso_mensual
1	51.22059	5.970588	43.92647	4340.118
2	29.95506	10.494382	43.76405	5513.292

En esta ocasión las variables más representativas analizando Sonora, son la edad y los años de estudio.

El primer clúster representado en Rojo en el lado derecho del gráfico se encuentra agrupando aquellos individuos que poseen en promedio una edad de 51 años con 6 años de estudio promedio, 44hs ocupadas e ingresos por $4340 moneda local.

Mientras que el segundo grupo color celeste ubicado en el lado izquierdo del gráfico posee un promedio de edad de 30 años quienes presentan en promedio estudios por más de 10 años, horas de ocupación similares al primer grupo e ingresos promedio de $5513 los cuales implican en promedio un aumento en salarios mensuales del 30% respecto al primer grupo.

Podríamos especular con que, en la muestra estudiada de Sonora, las personas catalogadas como más jóvenes poseen en promedio el doble de años de estudio y perciben consecuentemente un 30% más en ingresos que el grupo con edades promedio de 51años.

REGRESIÓN LINEAL MÚLTIPLE

Para la regresión lineal múltiple se emplearon las variables Ingresos, Años de escolaridad, Edad, Horas de ocupación.

Se tomó como variable dependiente a Ingresos dado que es la que queremos explicar y el resto de variables mencionadas como explicativas del fenómeno ingresos.

Antes de aplicar la regresión se llevó a cabo la detección y eliminación de datos atípicos para que no interfieran en el resultado final con interpretaciones erróneas.

Se empleó la herramienta de análisis de datos de Excel obteniéndose lo siguiente:

Estadísticas de la regresión	
Coeficiente de correlación múltiple	**0,852678793**
Coeficiente de determinación R^2	**0,727061124**
R^2 ajustado	**0,726910497**
Error típico	**4750,991529**
Observaciones	**10266**

Al estar ante regresión múltiple el R^2 mas significativo para las interpretaciones es el ajustado dado que penaliza el hecho de agregar más variables al modelo. En este sentido, podemos decir que las variables años de estudio, edad y horas de ocupación explican a la variable ingresos en un 73%.
Con respecto al desvío típico significa que al querer realizar estimaciones/predicciones tendremos un error en el resultado esperado de $\pm$ $4751.

ANÁLISIS DE VARIANZA

Descripción	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	3	6,1709E+11	2,05697E+11	9112,941829	0
Residuos	10263	2,31656E+11	22571920,51		
Total	10266	8,48745E+11			

Con el Valor F elevado en este caso con un valor de 9113, estaríamos en condiciones de decir que el modelo es adecuado, como también al observar el valor crítico de F igual a cero.

(Este estadístico F indica si todos los coeficientes de la regresión, conjuntamente, son distintos de cero. Es decir, indica si el coeficiente que acompaña a “Años de educación”, “edad” y el coeficiente que acompaña a “horas de ocupación” son distintos de cero, que es lo mismo que decir que son conjuntamente significativos. En general, queremos que el estadístico F sea lo más grande posible, o también que el “Valor crítico de F” sea lo más pequeño posible. En este caso, tenemos que el estadístico F elevado y su valor crítico es cero, por lo que sí podemos decir que los coeficientes son conjuntamente significativos.)

Descripción	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95,0%	Superior 95,0%
Intercepción	0							
edad	16,28	2,45	6,64	0,00	11,47	21,09	11,47	21,09
esc	491,57	9,00	54,61	0,00	473,93	509,21	473,93	509,21
hrsocup	45,27	2,35	19,30	0,00	40,67	49,87	40,67	49,87

Vemos que la variable que más repercute en el resultado final del ingreso mensual es los años de estudio (dado que posee el valor de su coeficiente varias veces superior frente al resto de variables explicativas), seguido de horas de ocupación y finalmente la edad del individuo.

En cuanto al **error típico**, el cual nos indica la variación de la estimación en los coeficientes posee valores acordes para buenas estimaciones en relación a los coeficientes.

A su vez los **estadísticos t** nos indican buena representatividad, como también la **probabilidad** o también conocido como **valor-p** se busca que los mismos sean cercanos a cero y en este caso tenemos los valores iguales a cero. Con respecto a los intervalos de confianza presentan un rango bajo por lo que son bastante aceptables.

Por todo lo anterior nos encontramos en condiciones de aceptar los resultados obtenidos por la regresión lineal múltiple como adecuados para realizar estimaciones.

La ecuación de regresión lineal múltiple quedaría como sigue:

$$Y = 16 * (edad) + 491{,}57 * (años\ de\ escolaridad) + 45{,}27 * (hs\ de\ ocupación)$$

APÉNDICE

El Scrip del trabajo se presenta a continuación

```
setwd("C:/Users/Martin/Desktop/R/Archivos/3.- GGPlot 2")

library(ggplot2)

library(readxl)

library(tidyr)

library(dplyr)

library(plotly)

library(ggplotly)

 enoe<-read_xlsx("mu_enoe.xlsx")

#MAPA DE PUNTOS DE DISPERSION INGRESOS VS AÑOS DE ESTUDIO VS TIPO DE EMPLEO

ggplot(data = enoe) +

   geom_point(mapping = aes(x = anios_esc, y = ingreso_mensual, color=tipo_empleo))

#CLASIFICACIÓN SEGÚN SEXO, TIPO DE EMPLEO, AÑOS DE ESCOLARIDAD E INGRESO MENSUAL

ggplotly(

ggplot(data=enoe)+

 geom_point(mapping = aes(x =anios_esc, y =ingreso_mensual))+

 facet_grid(tipo_empleo~sex))

#LINEA DE TENDENCIA EXPLONIENDO LOS AÑOS DE ESCOLARIDAD CON LOS INGRESOS MENSUALES

ggplot(data=enoe, mapping = aes(x =anios_esc, y =ingreso_mensual))+

 geom_point(mapping = aes(color=niv_edu), show.legend = TRUE)+

 geom_smooth()

 #PROPORCIÓN DE HOMBRES VS MUJERES EN LA MUESTRA ANALIZADA

ggplot(data=enoe)+

 geom_bar(mapping = aes(x=sex, y=..prop.., group=1, ))

#conteo de hombres y mujeres de la muestra

ggplot(data=enoe)+

 geom_bar(mapping = aes(x=sex, fill=sex))

#conteo de Individuos por sexo clasificado según su nivel de estudio

ggplot(data=enoe)+

 geom_bar(mapping = aes(x=sex, fill=niv_edu))

ggplot(data=enoe, mapping = aes(x=sex, fill=niv_edu,))+

 geom_bar( position = "dodge")+

 labs(title="Observaciones por sexo y nivel educativo", x="Sexo", y="-

Observaciones")

#idem anterior pero por poroporciones
```

```
ggplot(data=enoe, mapping = aes(x=sex, fill=niv_edu))+
  geom_bar( position = "fill")
library(ggplot2)
library(tidyselect)
library(htmltools)
library(plotly)
library(yaml)
#Generar un plotly basado en un boxplot
ggplotly(
  ggplot(enoe,aes(x=estado,y=anios_esc,fill=estado))+
    geom_boxplot(color='black',show.legend = T))
#ANÁLISIS DE CORRESPONDENCIA#####################################################################
library(ca)
library("ggplot2")
library("factoextra")
library("FactoMineR")
library("gridExtra")
 enoe<-read.delim("clipboard", dec=",")
 for (i in 1:ncol(enoe)) enoe[,i]=as.factor(enoe[,i])
str(enoe)
#Análisis descriptivo
ggplotly( ggplot(enoe,aes(x=estado))+ geom_bar(fill= "#DDB4EB"))
ggplotly(ggplot(enoe,aes(x=sex))+ geom_bar(fill= "#FFD4A5"))
ggplotly(ggplot(enoe,aes(x=edad))+ geom_bar(fill= "#41894A"))
ggplotly(ggplot(enoe,aes(x=pos_ocu))+ geom_bar(fill= "#43857A"))
ggplotly(ggplot(enoe,aes(x=ing_salarios))+ geom_bar(fill= "#FFEC75"))
ggplotly(ggplot(enoe,aes(x=niv_edu))+ geom_bar(fill= "#FFEC30"))
ggplotly( ggplot(enoe,aes(x=num_trabajos))+ geom_bar(fill= "#DDB4EB"))
ggplotly(ggplot(enoe,aes(x=tipo_empleo))+ geom_bar(fill= "#41894A"))
enoe <- MCA(enoe, graph = FALSE)
print(enoe)
print(enoe$var)
fviz_screeplot(enoe,geom="line")+
  theme_grey()
fviz_mca_var(enoe, col.var = "blue", addEllipses = FALSE, repel = TRUE) + theme_minimal()
#Contribuciones al eje 1
```

```
fviz_contrib(enoe, choice = "var", axes = 1, top = 15)+labs(title = "                Contribución de las Categorías para las Dimension 1")

#Contribuciones al eje 2

fviz_contrib(enoe, choice = "var", axes = 2, top = 15)+labs(title = "                Contribución de las Categorías para las Dimension 2")

#Contribuciones a ambos ejes

fviz_contrib(enoe, choice = "var", axes = 1:2, top = 15)+labs(title = "          Contribuciones de las Categorías para las Dimensiones 1-2")

#Gráfico de calor por contribuciones

fviz_mca_var(enoe, col.var = "contrib",

             gradient.cols = c("#00AFBB", "#E7B800","#FC4E07"),

             ggtheme = theme_grey()

             , repel = TRUE)


#boxplot

ggplot(enoe,aes(x="Descripción",y="Total"))+

  geom_boxplot(color='black',show.legend = F)

boxplot(enoe)

getwd()

library(cluster)

clusplot(wine.stand, k.means.fit$cluster,

         main='2D representation of the Cluster solution',

         color=TRUE, shade=TRUE,

         labels=2, lines=0)

# CLUSTER#############################################################

# Cargamos el data set

enoecluster<-read.delim("clipboard", dec=",")

for (i in 1:ncol(enoe)) enoe[,i]=as.factor(enoe[,i])

head(enoecluster)

str(enoecluster)

library(NbClust)

library(cluster)

library(ggplot2)

library(factoextra)

#detectar outliers

mean<-colMeans(enoecluster)

Sx<-cov(enoecluster)

D2<-mahalanobis(enoecluster, mean, Sx, inverted = FALSE)

print(D2)
```

```
pchisq(D2, df= 6, lower.tail= FALSE)

qchisq(.99, df= 6)

# Como existen diferencias en las escalas de las variables, debemos escalar las mismas

enoecluster.esc <- scale(enoecluster, center = TRUE, scale = TRUE)

#Obtenemos la matriz de distancias

dist <- hclust(d = dist(x = enoecluster.esc, method = "euclidean"), method = "complete")

res.dist <- get_dist(enoecluster.esc, stand = TRUE, method = "euclidean")

fviz_dist(res.dist, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))

matriz.dis.euclid<-dist(enoecluster.esc,method="euclidean",diag=TRUE)

round(print(matriz.dis.euclid),2)

#Analizamos el n?mero ?ptimo de clusters

library(ggplot2)

library(factoextra)

fviz_nbclust(x = enoecluster.esc, FUNcluster = kmeans, method = "wss", k.max = 15,

        diss = get_dist(enoecluster.esc, method = "euclidean"), nstart = 50)

fviz_nbclust(x = enoecluster.esc, FUNcluster = kmeans, method = "gap_stat", k.max = 15,

        diss = get_dist(enoecluster.esc, method = "euclidean"), nstart = 50)

fviz_nbclust(x = enoecluster.esc, FUNcluster = kmeans, method = "silhouette", k.max = 15,

        diss = get_dist(enoecluster.esc, method = "euclidean"), nstart = 50)

#Pedimos los indicadores para un análisis jerárquico con distancia euclídea y método de ward

library(NbClust)

res<-NbClust(enoecluster.esc, distance = "euclidean", min.nc=2, max.nc=8, method = "ward.D2", index = "alllong")

res$All.index

res$Best.nc

res$Best.partition

fviz_nbclust(res)

#Realizamos un dendograma

fviz_dend(x = dist, k = 6, cex = 0.6) +  geom_hline(yintercept = 4, linetype = "dashed") +  labs(title = "clustering", subtitle = "Distancia eucl?dea, K=4")

fviz_dend(dist, cex = 0.8, k=6,

        rect = TRUE,

        k_colors = "jco",

        rect_border = "jco",

        rect_fill = TRUE,

        horiz = FALSE)

#Probamos otro tipo de dendograma

library(igraph)
```

```
fviz_dend(dist, cex = 0.8, lwd = 0.8, k = 4,
          rect = TRUE,
          k_colors = "jco",
          rect_border = "jco",
          rect_fill = TRUE,
          type = "phylogenic")
fviz_dend(dist, cex = 0.8, lwd = 0.8, k = 4,
          rect = TRUE,
          k_colors = "jco",
          rect_border = "jco",
          rect_fill = TRUE,
          type = "circular")
# Seleccionamos al azar K Centroides en la base de datos normalizada.
# Para obtener los mismos resultados cuantas veces se reproduzca tenemos que
# usar una semilla de números aleatorios
# para K = 4 y asignaciones aletarioas nstart = 50
set.seed(123)
km.res <- kmeans(enoecluster.esc, 6, nstart = 25)
print(km.res)
fviz_cluster(km.res, data = enoecluster.esc, ellipse.type = "convex")
#Otra visualización
fviz_cluster(object = km.res, data = enoecluster.esc, show.clust.cent = TRUE,
             ellipse.type = "euclid", star.plot = TRUE, repel = TRUE) +
  labs(title = "Resultados clustering K-means") +
  theme_bw() +
  theme(legend.position = "none")
# Calculamos la media de cada variable por cluster utilizando los datos orginaleslibrary(dplyr)
A<-aggregate(enoecluster, by = list(km.res$cluster), FUN = mean)
print(A)
#Otra manera
library(kableExtra)
aggregate(enoecluster, by = list(cluster = km.res$cluster), mean) %>%
  kable(caption = "Media de Cada Variable por Conglomerado, Utilizando los datos originales" ,
        align = "c",
        digits = 6) %>%
  kable_classic_2(html_font = "sans-serif",
```

```
            lightable_options = c("hover", "striped")) %>%
 row_spec(0,
          bold = T,
          color = "white",
          background = "#219B6D")
# Validación interna de los clusters
library("clValid")
intern <- clValid(enoecluster.esc, nClust = 2:6,
                  clMethods = c("hierarchical","kmeans","pam",'clara'),
                  validation = "internal")
# Summary
summary(intern)
plot(intern)
library(psych)
anova1<-aov(Murder ~ km.res$cluster, data = enoecluster)
anova2<-aov(Assault ~ km.res$cluster, data = enoecluster)
anova3<-aov(UrbanPop ~ km.res$cluster, data = enoecluster)
anova4<-aov(Rape ~ km.res$cluster, data = enoecluster)
summary(anova1)
summary(anova2)
summary(anova3)
summary(anova4)
#Añadimos el número de grupo a los datos originales
enoeclustercluster <- cbind(enoecluster,km.res$cluster)
print(enoeclustercluster)
```

BIBLIOGRAFÍA

Aldás, J., & Uriel, E. (2017). *Análisis multivariante aplicado con R.* Madrid: Parafino SA.

García, J., Molina, J., & Bustamante, ä. (2018). *Ciencia de datos, Técnicas analíticas y aprendizaje estadístico en un enfoque práctico.* Bogotá: Alfaomega.

Hair, Anderson, Tatham, & Black. (1999). *Análisi multivariante.* Madrid: Prentice Hall, Quinta edición.

Lind, D., Marchal, W., & Wathen, S. (2012). *Estadística aplicada a los negocios y la economía 15° Edición.* México: Mc Graw Hill.